Biased Machines in the Realm of Politics

Dissertation submitted for the degree of
Doctor of Social Sciences (Dr. rer. soc.)

presented by

Simon Roth

at

Universität Konstanz

Faculty of Politics, Law and Economics
Department of Politics and Public Administration
Konstanz, 2022

Date of Defense
December 1, 2022

Supervisors
Prof. Dr. Susumu Shikano (Konstanz)
Prof. Dr. Karsten Donnay (Zürich)
Prof. Dr. Lyudmila Grigoryeva (Warwick)

Acknowledgements

This dissertation would not have been possible without the help and encouragement of many people. I've gotten a lot of support in a variety of ways.

First and foremost, I want to thank my supervisors. Susumu Shikano, my first supervisor, gave me the chance to come to Konstanz and study for another three years. I am particularly grateful to Susumu for fostering an atmosphere in which I was able to pursue riskier, long-term initiatives as well as my own research agenda. I am thankful for his often challenging but valuable feedback, funding, and general mentorship. These projects would not have been possible without his ongoing support. Second, I want to express my gratitude to my second supervisor, Karsten Donnay. Karsten made time for me often and always pushed me to improve my papers. Even after switching to Zürich, the assistance was plentiful and effective. Also, thanks to Lyudmila for being the third member of my dissertation committee, despite the fact that she departed Konstanz last year. Luckily, I had the opportunity to attend courses taught by all of my supervisors, where I learned a lot about Bayesian statistics, general computational techniques, and machine learning.

During my time in Konstanz the Graduate School of Decision Sciences provided me with a scholarship and research funds, allowing me to pursue a topic of interest for three years. The University of Konstanz's Research Committee (AFF) supplied further funds. I am grateful to these organizations and employers for their financial support with my study.

Aside from my supervisors and funders, I like to thank my co-author Sascha Göbel for his cooperation. His data collection on public voter registration files and Twitter statuses enabled the majority of the empirical research in the last paper. Moreover, my wonderful colleagues read, commented on, and helped to improve my dissertation projects. Many thanks to Fabio Votta and Benjamin Guinaudeau for reading the majority of my work, as well as for co-authoring one paper. Your intellectual inputs on most ideas were inspiring and I enjoyed the heated discussions about statistics. Furthermore, exchanges with other students, postdocs, and professors from the department and Graduate School of Decision Sciences (GSDS) enhanced this work. I've always appreciated the supportive environment at the Center for Data and Methods (CDM). I also want to thank everyone who came to all sort of colloquia. Thank you, Karin Becker and Jutta Obenland, for your exceptional administrative assistance in keeping me on schedule.

Finally, I'd like to express my gratitude to my family. Thank you to my parents and brother for their unwavering support. Most importantly, Anna, thank you. I am really thankful for your unwavering support, patience, motivating speeches, and more.

Summary

This dissertation addresses one of the most serious risks associated with automated decision-making: bias. This is not a new phenomenon, and decisions have always been biased, but automated decision-making multiplies the risks in many ways. The main challenges are: How can we detect biases? Who should be held accountable for biased predictions? And how can biases be mitigated or corrected? The three studies within this dissertation help answer these questions by emphasizing the importance of monitoring our own machine learning (ML) pipelines, auditing third party prediction systems, and exposing the potential abuse of predictive algorithms when given sensitive data.

The first paper (section 2) addresses the question of how to direct ML users to high-performing, robust, and fair models. ML systems have been shown to harm human lives via discrimination, distortion, exploitation, or misjudgment (Bandy 2021). Although bias is often associated with malicious behavior, this is not always the case. Inductive biases, for example, such as knowledge about parameter ranges or priors can help to stabilize a model optimization process. Furthermore, decomposing into statistical bias and variance, allows for model selection with minimum future risk. Since "all models are wrong, but some are useful" (Box 1976), we should analyze as many biases in ML as feasible before putting faith in our predictions.

The second paper (section 3) addresses the question about how to audit recommender bias on social media. The goal of this experiment is to quantify the causes of algorithmic filter bubbles by analyzing amplification bias in the recommender system of Twitter. Using simulation of human behavior with bots we can show that 'filter bubbles' exist and that they add an additional layer of bias to 'echo chambers'. More precisely, the algorithm responded far more strongly to bots that actually engage with content than to bots that just follow human accounts. This demonstrates that the Twitter algorithm significantly depends on human interactions to adapt to preferences of its users. This has serious consequences since users may be unaware of the large personalization bias that happens when they like or share content.

The third paper (section 4) addresses the question whether online communication is predictive of offline political behavior. We can predict the party affiliation and turnout likelihood of a person with fair accuracy using a unique dataset consisting of thousands of ordinary citizens, including their Twitter statuses, integrated with public US voter registration files. Our results show social media communication is sufficiently biased to provide information about attitudes and political behavior of an average person in the real world. We demonstrate how, in addition to us, political, commercial, or bad faith actors may acquire this sensitive data to build prediction models, for example, to influence a customers retail journey or perhaps worse discourage them from voting on scale.

Biases can limit the potential of ML for business and society by cultivating distrust and delivering distorting or discriminating results. However, if our societies can (1) implement effective data privacy regulations (2) require internal debaising steps and encourage external independent auditing (3) educate the broader public of biases and ways to report them (4) and invest in training interdisciplinary computational scientists, we may be better prepared for negative consequences of the next industrial revolution.

Zusammenfassung

Diese Dissertation befasst sich mit den wichtigsten Risiken im Zusammenhang mit automatisierter Entscheidungsfindung: "Bias" (Verzerrung). Dies ist kein neues Phänomen, und Entscheidungen waren schon immer verzerrt, aber die automatisierte Entscheidungsfindung vervielfacht das Risiko in vielerlei Hinsicht. Die größten Herausforderungen sind: Wie können wir Bias erkennen? Wer sollte für verzerrte Vorhersagen zur Rechenschaft gezogen werden? Und wie können Biases gemildert oder korrigiert werden? Die drei Studien in dieser Dissertation helfen bei der Beantwortung dieser Fragen, indem sie die Bedeutung der Überwachung unserer eigenen Prozesse für maschinelles Lernen (ML) betonen, Vorhersagesysteme von Drittanbietern prüfen und den potenziellen Missbrauch von Vorhersagealgorithmen bei der Bereitstellung sensibler Daten aufdecken.

Der erste Artikel (Abschnitt 2) befasst sich mit der Frage, wie ML-Benutzer zu leistungsstarken, robusten und fairen Modellen geführt werden können. Es hat sich gezeigt, dass ML-Systeme Menschenleben durch Diskriminierung, Verzerrung, Ausbeutung oder Fehleinschätzung schädigen (Bandy 2021). Obwohl Biases oft mit schädlichem Verhalten in Verbindung gebracht wird, ist dies nicht immer der Fall. Beispielsweise können induktive Biases wie das Wissen über Parameterbereiche oder Priors helfen, einen Modelloptimierungsprozess zu stabilisieren. Darüber hinaus ermöglicht die Zerlegung in statistischen Bias und Varianz eine Modellauswahl mit minimalem zukünftigen Risiko. Da „alle Modelle falsch, aber einige nützlich sind“ (Box 1976), sollten wir so viele Biases in ML wie möglich analysieren, bevor wir unseren Vorhersagen Glauben schenken.

Der zweite Artikel (Abschnitt 3) befasst sich mit der Frage, wie man Biases in Algorithmen auf sozialen Medien testen kann. Das Ziel dieses Experiments ist es, die Ursachen von algorithmischen Filterblasen zu quantifizieren, indem Bias auf Twitter analysiert wird. Durch die Simulation menschlichen Verhaltens mit Bots können wir zeigen, dass „Filterblasen“ existieren und dass sie den „Echokammern“ eine zusätzliche Verzerrungsebene hinzufügen. Genauer gesagt reagierte der Algorithmus viel stärker auf Bots, die tatsächlich mit Inhalten interagieren, als auf Bots, die nur menschlichen Konten folgen. Dies zeigt, dass der Twitter-Algorithmus erheblich von menschlichen Interaktionen abhängt, um sich an die Präferenzen seiner Nutzer anzupassen. Dies hat schwerwiegende Folgen, da sich die Benutzer möglicherweise der großen Personalisierungsverzerrung nicht bewusst sind, die auftritt, wenn sie Inhalte favorisieren oder teilen.

Der dritte Artikel (Abschnitt 4) befasst sich mit der Frage, ob Online- Kommunikation politisches Offline-Verhalten vorhersagt. Wir können die Parteizugehörigkeit und Wahlbeteiligung einer Person mit ziemlicher Genauigkeit vorhersagen, indem wir einen einzigartigen Datensatz verwenden, der aus Tausenden von normalen Bürgern besteht, einschließlich ihrer Tweets, die mit öffentlichen US-Wählerregistrierungsdateien integriert ist. Unsere Ergebnisse zeigen, dass die Kommunikation in sozialen Medien ausreichend verzerrt ist, um Informationen über Einstellungen und politisches Verhalten einer durchschnittlichen Person in der realen Welt vorherzusagen. Wir demonstrieren, wie neben uns auch politische, kommerzielle oder böswillige Akteure diese sensiblen Daten erwerben können, um beispielsweise Vorhersagemodelle zu erstellen, um das Kaufverhalten eines Kunden zu beeinflussen oder ihn vielleicht noch schlimmer davon abzuhalten Wählen zu gehen.

Biases können das Potenzial von ML für Wirtschaft und Gesellschaft einschränken, indem sie Misstrauen kultivieren und verzerrende oder diskriminierende Ergebnisse liefern. Wenn unsere Gesellschaften jedoch (1) wirksame Datenschutzbestimmungen umsetzen können (2) interne Schritte gegen Biases einfordern und externe unabhängige Audits fördern (3) die breitere Öffentlichkeit über Biases und Möglichkeiten, sie zu melden, aufklären (4) und in die Ausbildung interdisziplinärer Wissenschaftler investieren, sind wir möglicherweise besser auf die negativen Folgen der nächsten industriellen Revolution vorbereitet.

Contents

1 Introduction

The invention of steam engines revolutionized the way we work: it automated manual labor and craftsmanship that previously was tedious, time-consuming and expensive. Of course, prior to the invention, we already had the capacity to move along roads and produce a wide range of goods and services. However, when fuelled with coal, steam engines allowed people to move faster and produce more efficiently. The positive impact of the industrial revolution are many in terms of domestic production, the development of new technologies, income redistribution, the formation of political institutions. Yet, two centuries later climate change endangers the livelihoods of millions of people and we are starting to witness big detrimental consequences: the extinction of animal species, rising sea levels, devastating droughts and floods as well as all other disruptions of the fragile climate equilibria.

In many ways, machine learning (ML) is comparable to the steam engine. It enables the automation and acceleration of "data craftsmanship". Picking applicants for a job, distributing computer resources, selecting news material, interpreting documents, and identifying fraudulent behavior are all tasks that ML models can currently perform. These are not new tasks. Headhunters conduct manual searches for candidates to fill job vacancies. Translators used their language ability to translate texts. Like ML algorithms, headhunters and translators train themselves on previously known data to make a decision, for example, they ask themselves "who are good candidates for the job?", or "what words should the translated texts be composed of?". As for the steam engine, ML has the potential of triggering the next industrial revolution and place us on the verge of incredible societal changes. The potential realms of application are many, ranging from goods production, to medicine through entertainment. The democratization of ML has potentially a myriad of positive effects in relation to how products of the twenty-first century will be produced, how technology will be discovered, how wealth will be distributed or how political systems will be organized. Because fewer resources are necessary to bring an ML model into production and scale it up, ML is advancing considerably faster than steam engines. It does, however, just like the previous industrial revolutions, harbor the potential for long-term negative impacts. In contrast to climate change, which took several decades to be linked to industrial production, the negative effects of ML on society may already be evident right now.

Many voices are warning about the unrestricted roll-out of ML throughout society. Weizenbaum 1976 warned that bias might occur from both the data utilized in a program and the way the program is built. As a result, if the data used for training is biased, bias will be propagated by automated decisions. That means, if humans are less likely to choose people of color for management positions, algo-

rithms trained on prior hiring decisions will be less inclined to hire individuals of color. But algorithmic biases that affect minorities are not just limited to hiring decisions. For a variety of tasks ranging from image cropping, to face recognition, to credit and risk scoring, algorithms perform worse in regards to minority groups in the data (Garvie & Frankle 2016; Johnson & Fredrickson 2005). In 2020, the Association for Computing Machinery urged a halt to private and government use of facial-recognition technology due to clear bias based on ethnic, racial, gender, and other human characteristics, which harms the rights of individuals in specific demographic groups (Castelvecchi 2020). The misuse of ML can also be seen in various other fields. For example, numerous studies have found racial disparities in credit and insurance scoring in the US (Kennedy et al. 2011; of Governors of the Federal Reserve System (US) 2007). The outcomes for black Americans because of this bias are higher interest rates on home and auto loans with longer terms; increased debt collection default lawsuits, and an increase in the use of predatory lenders (Johnson 2010). Systematic discrimination against minorities using statistical systems is a reocccuring issue in the US. For example, the US court system uses predictive risk assessment to inform decisions on pretrial, bail out, early release, as well as sentencing (Angwin et al. 2016). The findings of research released in 2016 by ProPublica clearly reveals that estimates regarding non-reoffending defenders (from a recidivism prediction tool called COMPAS created by Northpointe Inc.) varied considerably for blacks and whites (Angwin et al. 2016). Black defenders were predicted to be twice as likely to reoffend as white defenders, but they also have a twice as high rate of false positives for basic recidivism (Angwin et al. 2016). The program makes the reverse error with whites: they are much more likely than blacks to be classified low-risk yet go on to commit further crimes. The software is still in use today by the US court system.

Today, most ML research is undertaken by the largest businesses in the world, including tech giants like Google, Facebook and Amazon. Some authors see the next stage in the evolution of capitalism - from the mass production of goods, service capitalism, financial capitalism, and now the exploitation of covert behavioral predictions generated from user monitoring (Zuboff 2015). The phrase "surveillance capitalism" refers to the practice of converting human experience into behavioral data and claiming social data as a free commodity (Cinnamon 2017). Although some of this data are used to enhance internal services, the majority are declared as a commercial surplus and put into ML in order to sell prediction products that influence the behavior of people through search, recommender, or advertising systems (Zuboff 2015). Feedback loops between users and the prediction systems result in online echo chambers and filter bubbles, which become fertile grounds for deep fakes, propaganda, COVID 19 disinformation, electoral interference and conspiracy

theories, among other things. The growing adoption and acceptance of ML-driven solutions plays into the hands of repressive regimes, expediting a global resurgence of authoritarianism, with China at the forefront of ML technology proliferation to authoritarian and illiberal regimes (Feldstein 2019). To combat not just the expansion of high-tech surveillance and repression worldwide, but also possible abuses at home, policymakers in democratic governments must understand the risks involved and consider ways to regulate its application (Feldstein 2019).

Political science is no exception to the enthusiasm around ML. Many publications already feature ML and even cutting-edge deep learning techniques. The democratization of programming, combined with lowered costs of computation, enable contemporary students in social sciences to train deep-learning models. Since the skills and understanding required to obtain an automated decision are not high, it is far more important to understand how to obtain accurate and unbiased automated decisions. This idea sets the ground for this dissertation. More precisely, this work attempts to clarify and organize the different risks at stake when training a ML model and releasing it for public consumption. As such, the first paper proposes a typology of biases that informs us about the risks we face when determining how to train a model. The second paper investigates the recommender system of Twitter and proposes an innovative research design to measure the ideological amplification bias of recommender systems. Finally, the third chapter carefully demonstrates that ML can be used to predict the political participation and orientation of American voters based on their everyday social media communication. While all of the papers share the objective of applying ML for text in political science, their epistemological contribution is threefold: The first paper serves an instructional role, addressing the questions how to direct ML users to high-performing, robust, and fair models. These abilities are required for applying ML to new substantive research questions in political science and beyond. In this sense, the second article employs ML to create a treatment policy for automated bots as well as to predict ideological biases using text as a measurement. The purpose was to find causal factors for algorithmic filter bubbles by auditing bias on the public-facing recommender system of Twitter. Within the third paper we predict party affiliation and voting behavior using ML based on a unique dataset consisting of thousands of ordinary citizens including their Twitter statuses merged with public voter registration files. We demonstrate the possibility that beside us - political, commercial, or bad faith actors can potentially access this sensitive data and predict political affiliation and other metrics from text, for example to manipulate citizens or discourage them to vote on scale.

The advancements in ML, including natural language processing (NLP), computer vision and other fields, open up new research frontiers for social scientists, including innovative measurements derived from data sources such as text, images

or videos (Banerjee 2017; Barberá et al. 2021; Beskow et al. 2020) and heterogeneous treatment effect estimation in causal inference (Grimmer et al. 2017; Künzel et al. 2019). However, the knowledge transfer from ML into various fields of science, which predominantly operate within inferential statistics, has created confusion in how ML models are supposed to be applied, how they are validated and how their output can be interpreted. As shown by the literature, simple goodness of fit measures are not enough to judge the predictive quality and fairness of a model. In this regard, the first paper introduces a typology for biases in ML predictions from discrimination, induction and error decomposition. First, algorithmic decisions can have a detrimental impact on human lives and thus require thorough fairness validation in order to address the vulnerability of protected groups of socioeconomic discrimination. Second, inductive biases are essential for a model to generalize beyond any specific dataset by learning biases inductively from data patterns. Design decisions for model families, parameters, or data transformations all contribute to inductive bias in the learning process, and we must be cognizant of our own contributions. Third, we may better understand model behavior and situations of under- and overfitting by decomposing the error into bias and variance. Low variance models are preferred because their outcomes are more predictable and less impacted by random chance. Following this typology, designing a model that is aligned with our objectives requires monitoring model biases along the entire computational pipeline, from data and development through validation and execution. Debiasing the data is the most common first step to adhere to several notions of fairness, which includes data balancing, feature removal or custom loss functions. Furthermore, techniques like sample splitting, bootstrap, or cross-validation can greatly improve our understanding the real risk of a model and error in unknown territories (Gong 1986; Vapnik 1991). Moreover, tuning models via cross validation greatly enhances robust model selection (Arlot & Lerasle 2016; Kohavi et al. 1995). Predictive models, when properly constructed and reviewed, have the ability to drastically increase both the effectiveness and equity of important decisions (Corbett-Davies & Goel 2018).

Over the past decades the public has been faced with many new prediction-powered services. Algorithm auditing as a research field is the response to this development and is employed to identify undesirable behavior in mostly proprietary algorithms. The primary focus of previous research has been on public search, advertising, and recommendation engines and the question how social biases manifest in technical systems. For example, authors have researched whether tailored advertising algorithms amplify discrimination (Crawford & Calo 2016; Lambrecht & Tucker 2019), they investigated the recommendation system of YouTube and if it favors extreme content (Faddoul et al. 2020), found that face recognition algorithms perform poorly on those with darker skin, (Raji & Buolamwini 2019) and examined

whether the Google search engine is biased toward particular news outlets (Lee & Riek 2018; Trielli & Diakopoulos 2019). As algorithmic systems grow more prevalent and powerful in society, these issues increase and with it the need to systematically understand its biases and impact on society.

With all of the prediction systems around us, businesses attempt to assist consumers in navigating the huge amount of content, such as websites, music, movies, people, and products, among other things. Most often, recommender systems are used which are a subfield of ML and, as such, are subject to similar bias concerns with some special caveats. Rather than being experimental, user behavior data is observational. The fundamental issue is that a user creates behaviors based on the exposed items, which confounds the observational data by self-selection and system-selection (Chen, Dong, et al. 2020). Exposure mechanisms allow consumers to choose certain items, content, or content providers which allows them to self-select into so-called echo chambers. However, since platforms incentivize content creators to produce a large amount of content, the volume of content that subscribers opted-in to see may still exceed any realistic consumption capacity. As a consequence, algorithmic recommendations are leveraged in addition to active user selection to limit content supply to a reasonable quantity. This is where the opportunity for filter bubbles arises, an algorithmic-induced state in which users are never exposed to viewpoints that the algorithms assume do not fit the personal profile of the user. Data generated with such a system is biased by several human and system confounders. The feedback loop for example, of exposure mechanisms and human behavior not only generates biases by sorting and ranking data, but also increases them over time, resulting in a compounding bias effect (Chen, Dong, et al. 2020). Even a small bias in the current state of a recommender system could be amplified over time which may lead to the emergence of discriminating or distortion biases if it is not properly dealt with.

The second paper investigates recommender bias and distortion through homogeneous political spaces. The two theoretical mechanisms we target are: self-selection into echo chambers and algorithmic selection into filter bubbles. This study provides important insights about how Twitter recommendation systems can distort the perceived reality of an user by showing less diverse and more personalized content. We focus on ideological bias and measure how the Twitter algorithm responds to ideological inputs, namely liking and retweeting of partisan content via automated social media bots. We find the causal driver for recommender bias on Twitter to be user engagement and interaction with content, which implies the system is distorting the timeline of its users to drive up engagement through personalization. Using an experimental research design and counter factual measurements of bias we observe an extra non-deliberate layer of recommender bias, aka filter bubble, which entails

normative implications.

Filter bubbles, may cause problems for fundamental democratic principles through the lens of liberal theories of democracy because opaque algorithmic filters that preselect content for their users limit their freedom to choose what kind of content they want to consume. Thereby violating fundamentally the freedom of choice (Bozdag & Van Den Hoven 2015). Proponents of deliberative democracy are likewise concerned, given the relatively low levels of viewpoint diversity and the small pool of accessible information, which makes it less likely that users would meet views that contradict their own (Bozdag & Van Den Hoven 2015). This results in situations in which not the best argument, but the most often heard one, prevails. Some authors emphasize the importance of informed interactions and evidence-based arguments in forming political ideas (Friedman 2014; Hardin 2009). Other democratic theories contend that public dialogue must incorporate and broaden the influence of (historically oppressed) minorities, marginalized groups, and portions of society with less means to campaign on their behalf (Parekh 2019; Sanders 1997; Young 2002). All these democratic principles are potentially at stake, if users receive a distorted version of reality that might by far of their stated preferences.

Political polarization is another growing concern in societies across the world. Issues for polarization are plenty from immigration, to climate change and lately vaccinations, among many others. People tend to interpret political information in a manner that confirms their prior beliefs contributing to rising political polarization. The biased assimilation of political information impedes efforts to persuade partisans toward positions of consensus and compromise. In the United States, for example, Democrats and Republicans have grown more ideologically divided in recent years, threatening both the social fabric and effective governance (McCright & Dunlap 2011). Scholars have frequently shown, particularly in the United States, that partisanship is a significant political identity and source of partisan biases (Greene 1999; Theodoridis 2017). A recent study shows that voters intentionally select into and identify with certain party groupings, and they respond more positively toward them than others (Bullock et al. 2013). Moreover, the identification with a political party is more and more associated with intolerance, hate, and mistrust of the opposing party (Iyengar & Westwood 2015; Mason 2015). Partisan bias influences not just how voters interact with one another, but also how they judge actions and hold opinions about facts (Bullock et al. 2013; McCright & Dunlap 2011). Electoral participation is also highly associated with political party affiliation. Voting enables the validation of a perceived party identification (Bolstad & Dinas 2017) as well as the promotion of favorable political results for the ingroup (Fowler & Kam 2007). Huddy et al. 2015 discovered that people who strongly identify with a party are also more likely to vote. The behavioral effects of political identification and

societal polarization should, by extension, drive voter participation, as well as be reflected in every day communication.

The last paper of this dissertation takes a look at how prevalent political biases are in everyday communication on Twitter and whether such biases in human communication are predictive of real world behavior, like voting for a certain party. We make use of a unique data set that combines US Voter Registrations with their respective Twitter data for thousands of regular US citizens. We predicted several outcomes using a neural network for texts, including party affiliation and election participation. The models for party affiliation were more predictive than those for electoral participation, suggesting that the language adopted was more biased. In addition, we extracted various features from the model predictions to have a better understanding of what the model had detected and reacted to. We can easily see how the system learns partisan biases from tweets with reasonable accuracy. We demonstrate the possibility that beside us - political, commercial, or bad faith actors can potentially access this sensitive data and predict political affiliation and other metrics from text, for example to manipulate citizens or discourage them to vote on scale.

The remainder of this dissertation is organized as follows. In the rest of the introduction, I put each of the papers into context and sketch out the contribution and outlook. The introduction is concluded with policy recommendations supported by the research within this dissertation. After that follow the research papers in section 2 to 4.

1.1 Biases in Machine Learning

In the first study of this dissertation I investigate biases in ML pipelines and demonstrate that before delegating any decision-making to an algorithm, a number of security risks must be addressed. Moreover, assessing whether each action is executed in a responsible way is out of scope due to increased autonomy and speed of ML systems. As a result, we must rigorously evaluate for common risk factors and biases, as well as identify viable counter strategies long before a model is believed valid. From the standpoint of both policymakers and practitioners, a more solid understanding of biases and their causes is important because it defines where we concentrate our efforts to mitigate the potential damage caused by biases. Social data and ML are transforming economies and societies with long-standing inequalities, and our models must avoid unwittingly perpetuating or worsening them (Zou & Schiebinger 2018).

Bias in a statistical process can be defined as the deviation of an expected value of some parameter or prediction from a benchmark value. A benchmark in statistics

could be the "true" population parameter or in ML the "ground truth" label of a past decision. Although the basic notion of bias is the same in both fields, their goal and process is vastly different. Traditionally, statistics is more heavily focused on inferences and ML on predictions. Making inferences about how a system operates in the real world and predicting what will happen next in the system are two distinctively important questions to answer (IJ 2018). Nonetheless, methods from both fields are used heavily in each community. For example, a very popular resampling method in ML is known as bootstrapping, which was originally designed to generate confidence intervals for statistical models with challenging variance properties to acquire analytically. In turn, predictive model validation can be used to enhance hypothesis testing in a statistical inference framework, ensuring model validity beyond a single sample of data. Moreover, ML approaches can address long-standing questions about the constraints of linear models, the criteria for assessing biases and discriminatory behavior, transparency in the context of discovery, and epistemological differences to inferential statistics.

1.1.1 State of the Art

The literature on biases in ML is plentiful but fragmented therefore I propose a typology to identify three dimensions that impact prediction quality: (1) discrimination, (2) induction, and (3) error decomposition. Whereas (1) is concerned with non-discriminatory requirements and formal fairness definitions, (2) explains the inductive design process by which a model is created, justified, and validated in ML, and (3) is concerned with the statistical bias and variance trade-off.

When dealing with human data or audiences, the idea of fairness is essential since any computation that does not adhere to this principle risks discrimination. Predictive systems have been demonstrated to impact human lives by discrimination, distortion, exploitation, or misjudgment (Bandy 2021). As ML systems are becoming more widespread in decision-making, the demand for fair algorithms will only increase, especially if biases happen on unlawful grounds. Hence, there are several formal fairness definitions in the literature to detect these biases from fairness through unawareness (Gajane & Pechenizkiy 2017), to individual fairness (Kusner et al. 2017), to demographic parity (Feldman et al. 2015), to equality of opportunity (Zafar et al. 2017) and finally counterfactual fairness (Kusner et al. 2017). Sometimes these fairness criteria are incompatible and more or less suited for certain situations (Kusner et al. 2017). When models are properly bias tested, predictive models have the potential to dramatically improve both the efficacy and equity of consequential decisions (Corbett-Davies & Goel 2018).

In ML inductive biases are the result of basic design choices that underpin the logic of every model. Mitchell 1980 refers to bias as a reason for preferring one

model realization over another that is not based on consistency with the in-sample error. The author asserts that there is no such thing as bias-free ML and that a productive bias is required. The widely held belief that computers, and hence ML, are objective or free of biases is wrong, because they are a fundamental requirement of any inductive learning system. Inductive biases are assumptions and design choices that must be fixed in order for any algorithm to derive logical conclusions from observable data. To arrive at effective design choices either the factual knowledge of the task domain should be sourced or methods used for automatically learning inductive biases (Baxter 2000). The initial choice of model and parameter space is often the most difficult challenge in any ML task; it must be broad enough to include a solution to the problem at hand, while small enough to enable effective generalisation from a limited number of samples (Mitchell 1980).

Statistical bias and variance are the two major components of error decomposition. In a nutshell, statistical bias of a learning system is the systematic error that the learning algorithm is expected to produce on average when fed training data (Dietterich & Kong 1995). Variance on the other hand measures the variability due to data sample differences or randomness in the statistical learning procedure. Therefore, the prediction error is made up of only two systematic components that can be controlled: bias and variance. A high variance model concentrates too much on the training data and fails to generalize to unseen data, hence it is overfitted and not repeatable (Belkin et al. 2019). Overfitting and underfitting are both undesirable conditions, and as a result, one key aim in algorithm design is to minimize statistical bias and variance independently (Belkin et al. 2019). Since statistical bias and variance are the outcome of the architecture of an algorithm, any modification to a ML pipeline might impact bias and variance, as well as overfitting or underfitting of a model.

1.1.2 Contribution

ML has grown at an exponential rate during the previous 20 years. Models continue to improve and become more user-friendly. The democratization of ML entails a surge in the usage of ML in almost every part of our lives, as well as the deployment of models by individuals who have not been qualified to build ML pipelines. This calls for the development of an uniform strategy for evaluating the predictive quality of ML models. According to research in ML, evaluating the predictive quality and fairness of a model based on its goodness of fit is insufficient. Hence, I propose a new typology of biases, distinguishing between discrimination, induction, and error decomposition. Finally, I demonstrate how those types of bias can impact each stage of a typical ML pipeline and mention possible mitigation strategies. The primary goal of this paper is educational, and it should be regarded as a practical guide to

bias in ML. By evaluating advances in ML validation, I intend to aid both researchers and practitioners new to the field of ML in becoming aware of biased predictions and advance our awareness of when ML models are high-performing, robust, and fair.

1.1.3 Outlook

Science does not truly advance via numerical predictions; rather, it progresses through mental or formal models that enable us to infer well beyond numbers. Statistical inference remains the backbone for meaningful conclusions drawn from data. However, ML predictions may be employed in scientific studies, such as for measurement or treatment purposes. This is exactly what the two papers that follow attempted to achieve. The second paper employs ML for texts to predict ideology from Twitter statuses, to give content-level treatments to our automated social media bots, and to measure recommender bias between chronological and algorithmic filtered timelines. The third paper uses text predictive modeling to identify partisan and electoral biases in exclusively citizen communication on Twitter, by using a unique Twitter database that has been merged with voter registration files from the US. Therefore it is necessary to build useful ML models that allow to perform valid statistical inferences on top of the prediction results.

1.2 Biases in Recommender Systems on Social Media

The second study in this dissertation, a team effort including Benjamin Guinaudeau and Fabio Votta is an auditing experiment to detect biases in social media recommender systems. We test two causal mechanisms in homogeneous political spaces: self-selection into echo chambers and algorithmic selection into filter bubbles. Echo chambers are a well-studied phenomenon that occurs when a person exclusively visits groups of friends with similar interests or constantly consumes the same news sources. This may limit the possibility for in-depth thinking on a subject, which increases the risk of naively adopting more radical ideas by only hearing "echoes of their own voices" (Sunstein 2017: 22). The idea of filter bubbles, on the other hand, indicates that a black-box algorithm ranks content in a user timeline, puts the user in a homogenous environment that may or may not be evident to the user, and is difficult to exit (Pariser 2011).

To help users navigate the massive amount of content, social media platforms have had to implement an exposure mechanism by which users can pick certain content or content creators (echo chambers) by becoming "friends" on Facebook, "following" on Twitter or Instagram, or "subscribing" on YouTube or Twitch. However, since platforms reward content creators to generate a large amount of content,

the opted-in volume may still surpass any practical consumption capacity. As a result, in addition to active user selection, social media platforms utilize algorithmic recommendation to restrict the content supply to a reasonable quantity (filter bubbles). Hence, the significance of recommender systems cannot be overstated, particularly at times when information overload is becoming more severe. Recommender systems may induce bias and systematically favor or punish specific content by sorting and ranking information. This may result in distorting biases emerging even when there is no intent at work. Because recommender systems are trained on previous human behavior or digital traces, they repeat human bias. This raises a slew of normative concerns about recommender systems' tendency to perpetuate biases.

1.2.1 Previous Research

Over the years the algorithmic auditing literature has come up with several ways to measure biases, most of them concentrating on public facing prediction systems (Sandvig et al. 2014). In an internal code audit, researchers obtain and analyze the code that makes up the algorithm which is not feasible for most proprietary systems (Weber & Kosterich 2018). Therefore, some authors use direct scraping data or systematic API queries of a system for auditing (Bono & Melo-Pfeifer 2011; M. H. Ribeiro et al. 2020). A step up in sophistication are sock puppets, where researchers collect data by creating automated bots to test the algorithm (Chakraborty & Ganguly 2018). Carrier puppets are similar to the sock puppet method, except that bots are designed to affect the real-world system and may "carry" effects onto end users (Ali et al. 2019). Finally, researchers collect data by hiring end users to test the algorithm via crowdsourcing (Bandy & Diakopoulos 2020). For this auditing study we settled on an sock puppet approach to treat the Twitter recommender system in order to simulate human behavior as authentic as possible on social media.

Some studies have investigated human behavior on social media using bots driving innovative research designs. Munger 2017, for example, did research that employed automated social media accounts to evaluate online activity on Twitter. Munger's research reveals that high-influence (e.g., many followers) "bots" acting as white male Twitter users are more effective than bots appearing as black males at reducing racial insults used by other white individuals on Twitter. Another study investigates the dynamics of information spread and concentration on Twitter using bots. Bots are used to test the reach of information by carrying out concerted efforts to distribute it (Mønsted et al. 2017). Another research employs neutral bots to investigate the impact of social media networks by investigating various biases arising from platform mechanisms vs user interactions (Chen, Pacheco, et al. 2021).

Researchers have, in the last few years heavily investigated, how mostly echo

chambers but also filter bubbles influence the attitudes of individuals and their political behavior. Studies conducted by Tewksbury 2003, Garrett 2009, and Beam 2014 confirm the assumption that users tend to access content that underpins their political attitudes. In a study that measures the effects of exposure to congruent partisan content, evidence suggests that exposure to supportive information increases affective polarization (R. K. Garrett et al. 2014) and polarization via homogeneous personal networks (Druckman et al. 2018). Others have found that the selective use of congruent media does not necessarily lead to the active avoidance of incongruous content (Garrett, Carnahan, et al. 2013). Several studies also suggest that the influence of algorithm-based platforms on people's opinion formation differs from common news assumptions. For example, a recent paper showed that exposing partisans to opposing views through bots on Twitter increased political polarization, at least for conservatives exposed to liberal viewpoints (Bail et al. 2018). De Graaf et al. 2012 show that endorsements of a given news source may trump the explicit partisanship affiliation of the source. Another study in cooperation with Facebook, hints at the conclusion that the effect of filter bubbles might be overestimated altogether (Bakshy et al. 2015). Further, as Beam is able to show in his study, users who are given the opportunity to explicitly tailor the recommendation mechanisms more frequently select articles with opposite attitudes (Beam 2014).

A review by Zuiderveen Borgesius et al. 2016 investigates the current empirical evidence of echo chamber and filter bubble effects on political opinions. While it appears that there are measurable and statistically significant effects of echo chambers on political attitudes of individuals, they remain rather low or moderate. The authors conclude that there is no empirical evidence that would warrant great concern about echo chambers. One of the more recent systematic assessments of the filter bubble hypothesis finds that content diversity among news stories selected by human editors and multiple set-ups of automatized recommendation systems (Möller et al. 2018). The authors conclude that both human and algorithmic methods of prioritization provided a good diversity of opinion. There is also debate in the literature on whether echo chambers are a problem at all. For example, Guess et al. 2018 describe the somewhat limited evidence of echo chambers online and even talk about an echo chamber about echo chambers.

In the past the echo chamber phenomenon has failed to materialize in some studies examining Google News (Haim et al. 2018), Facebook News Feed (Bechmann & Nielbo 2018), and Trending Stories in Apple News (Bandy & Diakopoulos 2020). However, there are other studies finding evidence for recommender bias and distortion via filter bubbles. For example, research looked at whether the recommender algorithms on three social media platforms, YouTube, Reddit, and Gab, encourage extremist content (Reed et al. 2019). They performed the analysis using

a small number of bots and discovered that especially YouTube prioritizes extreme right-wing videos following contact with comparable content. They conclude that these algorithms may favor extreme information over more moderate viewpoints. A recent auditing study conducted by Twitter researchers, using a massive-scale experiment involving millions of Twitter users in several countries, finds that the political right enjoys higher amplification compared to the political left (Huszár et al. 2022). The purpose was to better understand the amplification of elected officials' political content on the algorithmically ranked Home timeline versus the reverse chronological Home timeline. The echo chamber and filter bubble mechanisms are featured prominently in the literature. Yet empirical evidence is often either contradictory or inconclusive in particular regarding the role of ML systems vs user behavior. The fact that both concepts are often used interchangeably in the literature lead us to include both mechanisms in our experimental auditing approach.

1.2.2 Contribution

The second paper proposes an experimental method for systematically testing the biases of social media recommender systems. We focused on ideological biases in the run-up to the 2020 US Presidential Election and studied how the Twitter algorithm reacted to ideological inputs other than directly following content creators. We can demonstrate that 'filter bubbles' exist and that they add an extra layer of bias to 'echo chambers' on empirical grounds. Only the explicit stimuli had a significant impact as compared to the chronological timeline and the control group (following). When bots merely follow congruent accounts and do not engage with the content, the percentage of crosscutting content in the recommended timeline is identical to that in the chronological one, indicating no recommender bias. In contrast, we may see a substantial bias when it comes to implicit cues (liking and retweeting). Even if a bot followed as many conservative as liberal accounts, through liking or retweeting the algorithm decreases the percentage of cross-cutting content by 6% on average after only 3 weeks on the platform. More specifically, the system reacted significantly more strongly to bots actually engaging with the content than to bots that just followed human accounts. This shows that when it comes to understanding its users' preferences, the Twitter algorithm relies heavily on user interactions. This has significant ramifications since users may be unaware of the substantial personalisation bias that occurs when they like or share content. In sum, the degree of recommender bias or filter bubble a user is exposed to changes as a function of the user inputs supplied.

This study makes several methodological contributions. To safeguard our conclusions, we first pre registered an experimental research design well in advance of the project's start and compiled a public pre-analysis plan. Second, the majority of

studies utilizing automated social media bots to study recommender bias are either statistically underpowered, lack an experimental research design, or a counterfactual measurement of bias. In comparison, with 170 bots, thousands of online sessions, and millions of Tweets, our data sample is large enough to derive accurate inferences from the data and models. Furthermore, the reconstructed chronological timeline of all content providers at the time of each bot's online session naturally provides a counterfactual benchmark to study recommender biases. By comparing each bot's ideological score to its chronological benchmark, we can quantify recommender bias and the causal impact of user inputs on this bias. Finally, bot treatments are based on a neural network text model trained on tweets to predict the ideological identification of US Members of Congress. Using cross validation, we chose a robust model to be utilized throughout the experiment to determine whether to interact or not depending on the predicted ideology score at tweet level. Therefore, we used ML predictions as a measurement of ideological recommender bias as well as a treatment policy.

1.2.3 Outlook

Using bots for auditing recommender biases offers various advantages, including the elimination of human confounding and the collecting of huge amounts of data on a minimal budget. When compared to human surveys, this method is less expensive but takes longer to complete since each bot had to be constructed separately utilizing several phones for sim card verification. To summarize, it took a significant amount of effort to put up an operational bot infrastructure with fake user accounts, VPN rotation, and captcha solving. Throughout the research process, we realized that a variety of treatments, such as establishing that recommender bias is not only ideological, but rather a personalization bias, would have been intriguing. Furthermore, it would be interesting to investigate if we may find reversal effects if the bots start to engage with opposing ideological content more often. This would reflect how simple or difficult it is to exit an algorithmic filter bubble. Another interesting question could be tackled with a dose-response approach, where one can change the degree to which an automated account interacts with their own and opposing political sides, therefore exploring how diverse and homogeneous patterns of behavior impact the algorithm in a continuous space. All of these issues were beyond the scope of our study, which aimed to propose a simple and robust experimental research design for detecting recommender biases, aka filter bubbles, and determine their causes. Audits that concentrate on public-facing advertising, search, or recommender systems are particularly critical given the rising number of opaque algorithmic technologies that are already implemented. The lack of a FDA for algorithms (Tutt 2017), allows researchers to continue to undertake audits to explain the safety of public-facing al-

gorithms and assist in determining if and how they serve the public good. We think these audits should continue because they give important insights into how algorithms bias the social media ecosystem and impact its users. Future audits may benefit from looking at use cases, technologies, and domains other than Twitter, Facebook and Google. Moreover, a dedicated bias testing framework, provided by social media companies, would facilitate auditing and transparency of results, as well as would encourage replication of previous studies.

1.3 Biases in Language on Social Media

The goal of the third study, a collaboration with Sascha Göbel, is to test whether we can use citizen social media communication to predict their political behavior like voting for a certain party. We have access to a unique data collection that combines US Voter Registrations with their respective Twitter data for thousands of regular US citizens. We predicted several outcomes using a neural network for texts, including party identification and election participation. Our data and method offer insights on whether (exclusively) citizens complete and everyday communication bias on Twitter is predictive of their individual partisan affiliation and indicative for electoral participation. We demonstrate the possibility that beside us - political, commercial, or bad faith actors can potentially access this sensitive data and predict political affiliation and other metrics from text, for example to manipulate citizens or discourage them to vote on scale (see Cambridge Analytica).

1.3.1 Previous Research

Individual party preferences and vote choice are highly influenced by interpersonal communication, according to the literature on political communication (Beck et al. 2002; Huckfeldt & Sprague 1991; Kenny 1998; Partheymüller & Schmitt-Beck 2012). Similarly, multiple studies have shown that interpersonal communication has a positive impact on voter participation (Carlson et al. 2020; Eveland & Hively 2009; McClurg 2006). However, there is a paucity of empirical data on whether interpersonal communication genuinely transmits such information. Existing research, which are mostly survey-based, lack respondents' unique communication records (Schmitt-Beck & Partheymüller 2016). As a result, rather than being deduced from real communications, communicated information is often presumed based on personal attributes. The effects of interpersonal communication are measured by the structure of a person within a network and the frequency of activities. As a result, real communication and its substance have taken a back seat in the extant literature (Schmitt-Beck & Partheymüller 2016). Interpersonal communication is abundant on social networks. In this sense, computer scientists have intensively investigated this

source. However, this rather technical work focuses on self-reported political orientation, is usually based on political rather than ordinary talk, confounds elite and citizen communication, and concentrates on specific vocabularies or linguistic features (Hinds & Joinson 2018).

The study of political language is a branch of literature that examines actual communication in relation to political attributes. This body of literature is mostly concerned with detecting language disparities along political ideologies. The majority of the research focuses on political elites. Cichocka et al. 2016 and Diermeier et al. 2011, for example, look at disparities in language use between liberal and conservative politicians, while Slapin & Kirkland 2020 look at pronoun use, sentiment, and complexity of speeches from politicians that disagree with the party line. Scholars have only lately undertaken such research with regard to citizens. Okdie & Rempala 2019 and Sterling et al. 2020, for example, investigate language disparities between liberal and conservative undergraduate psychology students and Twitter users across numerous linguistic categories. Although these studies have made significant contributions to our knowledge of political language usage, they are constrained by an exclusive emphasis on political orientation and a narrow view on communication that often concentrates on psychological factors. Analyses are often limited to particular linguistic traits, which are commonly detected using the Linguistic Inquiry and Word Count dictionaries.

Voter preferences and participation are seen as outcomes that may be predicted from behavioral trace data in the literature on political behavior. Barberá 2015 measures social media users' party allegiance using Twitter follower networks, while Kosinski et al. 2013 utilizes Facebook likes. Bach et al. 2020 uses online traffic information from survey respondents to predict their political opinions and election participation. Beauchamp 2017 aggregates text messages from Twitter to estimate voter intentions at the state level. Individual communication on social media has gotten a lot of interest in computer science. However, the emphasis in this literature is less on the message and information provided via communication. It focuses on technical elements in order to find the best performing text classifier for political ideology. As a consequence, these studies focused on political communication rather than general communication, blended communication between political elites and citizens, and used restricted vocabulary or specific language traits (Hinds & Joinson 2018).

1.3.2 Contribution

This article contributes to the current literature by examining whether the whole content of entirely citizen communication reveals their party preferences and election participation. We offer a theory about why communication should reflect voter

preferences and involvement in the first place to inspire and frame our empirical investigation. We contend that the group norms that comprise the resultant political identities have an impact on communicative behavior (Hogg & Reid 2006). Consequently, individuals' perceived identities should be reflected in communication and represent their partisan bias and involvement in politics. We discover that communication, to varied degrees, conveys voter preferences and involvement. Individual communication is strongly predictive of party preferences, implying more partisan bias in day to day communication on Twitter. Communication is also moderately predictive of voter turnout, but only in low-stimulus elections. This implies that communication conveys information about the voting behaviors of potential voters but not about the electorate as a whole.

We use a new dataset that combines individual-level administrative information with digitally recorded communication for over 36,000 voters to overcome data constraints in previous studies. We use data from Göbel 2020, which combines public voter registration data from the state of Florida with public Twitter accounts, and which we extended with data from five more US states for further validation. Based on voters' official party registration and verified turnout in the 2018 general and primary elections, this data enables us to predict both our outcomes of interest, partisan affiliation, and vote participation. Furthermore, over 11 million text messages were collected, which provides us with direct access to communication content at the message level rather than its source. The next step was to create an artificial neural network for natural language and text data (Chollet & Allaire 2013). This method enables us to examine the whole spectrum of communication, including a vocabulary of over 600,000 words, as well as the semantic fingerprint capturing word context, location, and order. We use repeated K-fold cross-validation to train and assess our model. This allows us to examine the extent and variability by which communication usually predicts our outcomes of interest. Furthermore, we use Local Interpretable Model-agnostic Explanations (Ribeiro et al. 2016b) to provide substantial insights into what citizens are saying that corresponds to their partisan bias and electoral engagement.

1.3.3 Outlook

This study has several limitations that may be taken into account in future research. First, our conclusions are based on Twitter communication. However, Twitter is only one of numerous social networks, and the findings may vary for others. More crucially, we don't know how well the content of Twitter chats generalizes to real-world interactions, or if our results apply there as well. Second, data constraints compel us to omit unregistered voters from our analysis, despite the fact that many of them have strong party affiliations and are not chronic nonvoters. Registered

voters, as well as Twitter users, may be more engaged in political discourse than the general population. Third, we concentrate on the United States, where political identities are extensively formed. Our discovery that predictive signals are based on political identities raises the issue of whether interpersonal communication transmits electoral characteristics in circumstances where such identities are less established. Finally, we hoped to make publicly available pre trained neural networks for ideology prediction based on tweets. However, we had to acknowledge that, respecting data privacy and the possible abuse of these models, the potential benefit did not outweigh the risk of damage.

1.4 Conclusion

Machine learning (ML) is already driving the next industrial revolution, and its democratization has the potential to positively alter the lives of billions of people. It already affects how goods are produced; new and recent discoveries could not have happened without ML. The use of ML algorithms in contexts of market competition also affects how wealth is redistributed among human beings. ML is also used for political purposes and has the potential of deeply altering how democracies work. The automation of analytical thinking poses several challenges to our current society. For example, if labor (both employees and managers) is automatised, it cannot be the main factor of wealth redistribution: how can we manage to redistribute wealth in a society where production is automatized? How can we train people to make the most of automated tools? How can we control and correct racial and gender biases? How can the state regulate virtual reality, where coercion is much harder than in real life? Answering these important questions will require a lot of multidisciplinary research in the coming decades.

This dissertation addresses an issue associated with any automated system of decision-making: bias. This is not a new phenomenon. All decisions, both human-based and machine-based, suffer from bias. However, after working decades on developing protocols and rules to protect minorities from biased human-based decisions, the raise of ML endangers many of the advances made in the last century. Moving from human to machine decisions potentially multiplies the risk of biases in at least three ways. First, harmful biases are ambiguous and difficult to detect if not systematically investigated. Second, the lack of transparency about human input and machine output bias makes enforcing responsibility difficult. Third, the vast speed, quantity, and autonomy of ML systems complicates bias reduction. The following challenges emerge: How can we detect biases? Who should be held accountable for biased predictions? And how can biases be mitigated or corrected? There are no quick solutions for these broad and complex problems, but the stud-

ies undertaken in the context of this dissertation provide four elements of answers, setting us on the path of finding a stable and fair solution to the problem of ML bias.

1. Regulation. Algorithms are powered by data. Without data they are just a set of rules and parameters. As a result, ML regulation is closely related to data privacy legislation. As shown in paper 3, publicly available data may be simply exploited to predict and perhaps manipulate political behavior. This is only possible due the fractured and weak approach to data privacy in the United States. Instead of enacting a single comprehensive piece of legislation, as the EU has done, the US has chosen to create sector-specific privacy and data protection regulations that operate in tandem with state laws. While US legislation addresses data security and the importance of private records, privacy is frequently absent from the discussion or aligned with deregulatory approaches to privacy. They often argue that data collection is inevitable and that regulation should instead focus on use and service provider accountability (Coos 2018). With the General Data Protection Regulation (GDPR), the European Union has taken a new approach to data protection, restricting who gets access to specific categories of data. GDPR introduces privacy by default and by design, stricter controls over cross-border data transfers and cements EU citizens' right to reassure, essentially allowing them to request the deletion of their data (Coos 2018). The European data privacy protection is therefore an important pillar in the current ML security framework. It allows to better control the risk and biases in ML by preventing sensitive data being collected or distributed in the first place.

2. Mitigation. Without research biases are not recognized and cannot be mitigated. Therefore, addressing biases properly requires an institutional framework that allows experts to independently test and oversee algorithms before they go into production. As shown in paper 1, ML debiasing frameworks may assist in detecting code bugs and biases and increasing confidence in the performance of ML systems. Debiasing frameworks, on the other hand, favor service providers in determining what constitutes discrimination or malicious behavior, when it happens, and when it is adequately handled. While debiasing frameworks have been invented with the best of intentions in practice, they are likely to reinforce the power of service providers. They decide in practice what are acceptable social harms, while also weakening regulatory protections for example by normalizing the collection of sensitive data, or by certifying that systems are bias-free (Balayn & Gürses 2018). As established in paper 2, the recommender system of Twitter produces considerable bias, which they either accept or are unaware of. As a result, independent researchers must evaluate biases in proprietary or public-facing prediction services in order to objectively assess the risk of biases in commercial ML systems.

3. Education. Raising awareness regarding the causes and forms of existing biases is a critical step in countering biases. ML biases play a role in credit denial based on race, underestimating high-risk healthcare requirements for people of color, and prioritizing hiring based on gender. Each case of ML bias is complicated, demanding of investigation and focus on the plethora of factors at play in each setting. As a result, efforts to increase public awareness of biased prediction algorithms are just as vital as providing a method to appeal the automatic decision with a human in the loop. These problems often need more than technological solutions, such as evaluating whether a system is fair enough to be exposed to the public and in which instances entirely automated decision making should be authorized at all.

4. Training. ML Practitioners must be trained about the danger of biases, about the many forms they can take, and about which strategies can mitigate them. The causes of biases cannot be discovered without thorough knowledge of how algorithms are designed, taught, and deployed. Conceptualizing bias, on the other hand, requires the capacity to consider measurement errors, messy data sources, and behavioral mechanisms. Because biases adapt to their environment, detecting them needs analytical thinking and cannot be fully automated. Computational social scientists play a key-role in confronting the issues of ML biases as it requires both the analytical thinking taught in social sciences and the technical abilities taught in computer sciences.

Biases can limit the potential of machine learning for business and society by cultivating distrust and delivering distorting or discriminating results. However, if our societies can (1) implement effective data privacy regulations (2) require internal debiaising frameworks and encourage external independent auditing (3) educate the broader public of biases and ways to report them (4) and invest in training interdisciplinary computational scientists, we may be better prepared for negative consequences of the next industrial revolution.

2 Bias in Machine Learning: Curse and Blessing

Simon Roth

Abstract

The internet, big data, and statistics have had a profound impact on many disciplines of study and aspects of life, and machine learning (ML) is no exception. ML offers enormous promise in the social sciences, such as massive data discovery, measurement, and treatment prediction. However, before transferring any responsibility to a decision-making algorithm, regardless of the context of an ML solution, a number of security risks must be addressed, the majority of which are related to bias. As non-ML professionals increasingly use ML, this article discusses biases in ML and showcases solutions for coping with the most prevalent ones in practice. As shown by the literature, the simple goodness of fit is not enough to judge the predictive quality and fairness of a specific model. In this paper, I proposed a new typology of biases in ML and distinguish between discrimination, induction and error decomposition. Finally, I show how those type of biases matter at each step of a classical ML pipeline and mention possible mitigation strategies. The main objective of this paper is educational and should be seen as a practical guide to bias in ML.

2.1 Introduction

The rising automation in our lives is being driven by machine learning (ML). Every smart service, from recommender systems, to fraud detection or risk assessment, leverages ML in order to provide prediction and value on demand. All these algorithms are driven by data, which has become the most precious commodity in this day and age of the internet. Despite the fact that the amount of data generated on a daily basis is always expanding, our comprehensive knowledge of it has trailed. ML models were created in response to the pressing need to automate the process of information discovery (Cunningham 1995). ML is now not just the gold standard for big data discovery, but also for the majority of prediction tasks. However, before delegating any decision-making to an algorithm, a number of security risks must be addressed. Any optimisation can produce bad outcomes, inadvertently and unintentionally causing harm via biased predictions. Biased models can have a severe impact on people if they are used to inform consequential decisions in banking, criminal justice, health, and other fields (Barocas & Selbst 2016; Berk 2012; Chouldechova et al. 2018).

Bias in a statistical process is defined as the deviation of an expected value of some parameter or prediction from a benchmark value (Breitung & Knüppel 2021). A benchmark in statistics could be the "true" population parameter or in ML the "ground truth" label of a past decision. Although the basic notion of bias is the same in both fields, their goal and process is vastly different. Traditionally, statistics is more heavily focused on inferences and ML on predictions (IJ 2018). Making inferences about how a system operates in the real world and predicting what will happen next in the system are distinctively important questions to answer. To better understand the difference between statistics and ML I will dive into their distinct goals and processes namely inference and prediction.

From a programming point of view, models keep getting better and easier to deploy. The democratization of ML inevitably implies an increased use of ML in most aspects of our life as well as the deployment of models by people, who were not trained to develop ML models. This calls for a standardized strategy to evaluate the prediction quality of ML models. As shown by the literature, the simple goodness of fit is not enough to judge the predictive quality and fairness of a specific model. In this paper, I propose a new typology of prediction biases and distinguish between between discrimination, induction and error decomposition. Finally, I show how those types of biases matter at each step of a classical ML pipeline and mention possible mitigation strategies. As such, it is important to reevaluate past research on biases in ML and conduct experiments, to identify best practices moving forward. The main objective of this paper is educational and should be seen as a practical

guide to bias in ML.

2.2 From Inference to Prediction

In the study of systems with data, two of the most important goals are inference and prediction (IJ 2018). Statistics is primarily concerned with inference, which is the act of developing a mathematical model of the data-generation process in order to formalize information or test a hypothesis about how a system operates (IJ 2018). Machine learning (ML), on the other hand, is concerned with predicting unknown outcomes or future behavior of a given system based on past behavior (IJ 2018). This applies to supervised and unsupervised models, which are both used for predictions. Despite the fact that both fields statistics and ML deploy statistical models such as linear or logistic regression, for example, to quantify gender result disparities or predict consumer demand, their goals are most often not the same. Whereas statistical inference is used in the scientific process in order to derive evidence for theories, ML predictions are more commonly used in business use cases with decisions on demand.

2.2.1 Statistical Inference

Science mainly progresses through statistical inference that implies a theoretical framework or hypotheses, as well as a statistical model for empirical testing (Dangeti 2017). Thereby statistics is usually used to calculate population inferences from a sample of that data by fitting a model, which provides confidence boundaries (frequentist) or credible intervals of the posterior (Bayesian) in which a parameter of interest is likely to be bound (Shikano 2019). It allows us to compute a measurable degree of uncertainty/confidence that a discovered correlation between two or more variables reflects a 'real' effect that is unlikely to be caused by random chance (IJ 2018). Despite the fact that the quantity of interest is not directly the prediction $\hat{y}_i$ but the parameters $\hat{\beta}_k$, statistical inference is also providing some form of "predictions" (Mullainathan & Spiess 2017). For example, in a linear model $\hat{\beta}$ is just a compression of $\hat{y}$ as we estimate $\hat{y}_i = \hat{\beta} x_i$, that leads to $\hat{\beta} = (x^{'} x)^{-1} x^{'} y$. In this case, we interpret a compressed prediction in the form of a parameter, to make inferences and extrapolate knowledge from a sample of data to a population $(N + 1)$. In other words, statistics offers abstracted evidence and "predictions" regarding a larger population or system, which is often used for policy recommendations and the expansion of the scientific knowledge base (Mullainathan & Spiess 2017).

Causal inference is the gold standard in statistical inference, which requires additional strong assumptions unless we use randomized treatment assignments. Causality arises from the fact that it is about active intervention $P(Y|$ set $X = x)$, rather

than passive observation $P(Y|X = x)$ as it is for ML (Wasserman 2015). The difference between both is the fact that correlation is not causation. This statement can be noted as $P(Y|X = x) \neq P(Y|$ set $X = x)$. The average treatment or causal effect $\theta = E(Y_1) - E(Y_0) = E(Y|$ set $X = 1) - E(Y|$ set $X = 0)$ is the mean response if everyone was compelled to take the treatment minus the mean response if everyone was compelled not to take the treatment (Wasserman 2015) . If the treatment is given at random, correlation equals causation (Wasserman 2015). But in certain situations, a randomized experiment is not possible or not desirable. For example, it would be unethical to study the impact of smoking on lung cancer with a randomized control trial given the harm this would do to participants who are chosen to smoke. We may also not be able to study wage impacts after marriage in randomized experiments since we cannot randomize who is going to get married or not. Statistical inference is the process of discovering connections between variables and determining the significance of such associations even when no randomized trial data is available.

The justification for a statistical inference model depends on whether we believe it correctly represents the underlying relationships of the system. Because statistics was created to test hypotheses, researchers typically have a set of theoretical assumptions about how a particular dataset was generated, and this information has been encoded into a model. Essentially, we want to know whether our theoretical model matches the empirical model between the input and output variables, and how distant this data realization is from random chance. Integrating deductive information about the system into our design, measurement, and statistical model of that system requires considerable work. This deductive process is referred to as statistical inference, and with randomized trial data, it is referred to as causal inference. For statistical inference we usually do not see a training or a test set since we are not attempting to minimize empirical or future risk. This model may still be used to generate "predictions" but it will be assessed without a test set, concentrating on the relevance and robustness of the model parameters (Shikano 2019). Various methods, including uncertainty tests, confidence and credible intervals, bootstrap, and others, help in establishing a valid model and inferences from that model.

Using statistical models for hypothesis testing entails a lot of inferential assumptions that are not required for the same model used for predictions in ML. For example, linear regression, fits a line (or hyperplane) that minimizes errors across all data points, assumes the data is linearly correlated plus some random error, which is often Gaussian in nature. For valid statistical inferences from such a model the following data and model assumptions should not be violated (Gelman & Hill 2006; Wasserman 2015): (1) the data is independent and fixed, (2) the linear model is sufficient, (3) the variance is constant, (4) the errors are normally distributed, (5)

a small number of model parameters, and (6) low feature multicollinearity. These assumptions allow us to assert that the least squares estimates of the parameters β have the smallest variance among all linear unbiased estimates according to the Gauss-Markov theorem (Hastie et al. 2009). However, when observations are neither independent nor identically distributed (iid), the first assumption makes inference rather difficult. The iid assumption is a relic from a time when regression was only used in natural experiments in which the data was really fixed and endlessly repeatable (Wasserman 2015). Linearity and constant variance are both unlikely to hold up in practice and the regression function does need to be close to linear in reality (Wasserman 2015). When high collinearity is the norm rather than the exception, particularly in high-dimensional scenarios, model assumptions are dubious at best (Wasserman 2015). In sum, statistics requires simple data structures and models that come with simplifying assumptions about a system, that need to hold up for valid inferences and conclusions from a model.

2.2.2 Models from Statistics and Machine Learning

Statistics has created a series of mathematical models, such as linear regression and principal component analysis, that are extensively used in both the statistics and ML community. A statistical model is either used to infer something about the relationships within the data or to be able to predict future values, which often goes hand-in-hand. Statistical models were created to handle small to medium-sized data sets and require human selection of the feature space and model parameters. That is, we usually specify which variables to include and what prior we assume for the probability distribution of the outcome (Dangeti 2017). Although statistical advancements like regularization decrease the feature space automatically with a shrinkage penalty (ridge and lasso regression), statistical models typically emphasize interpretability above performance (Tibshirani 1996). The term interpretability refers to models that are transparent in terms of how outcomes are produced and where a user can comprehend the underlying decision logic of model (Vokinger et al. 2021). For high-risk or high-stakes prediction tasks, it may be good advice to use simple and interpretable approaches, which are typically statistical models in order to better justify your decisions (Vokinger et al. 2021).

In contrast, ML models work mostly without assumptions about data or outcome distributions but rely on assumption free inference through sample splitting. ML models can provide useful predictions, even when the data is non-linearly related or when the data has hundreds of features or millions of observations. Their main focus is predictive performance and often interpretability is traded off for black box predictions (Vokinger et al. 2021). The inner workings of most modern ML algorithms are too complex to humanely interpret and therefore require extra efforts to

get meaningful information about states, bugs and biases of a model. Understanding how an ML model behaves allows developers to evaluate the insights against past information from research to guarantee that a model takes into account known risk factors. Explainability in contrast to interpretability implies that a second model is developed to describe the first model, which is not necessarily transparent (e.g., because it can have millions of parameters). However, care is advised since ML explainability may be incorrect or meaningless due to other underlying second model assumptions. Explainability in ML is supported via various software tools, for example SHAP uses a game-theory (Lundberg & S.-I. Lee 2017) and LIME a local interpretable approach to explain black box predictions (Ribeiro et al. 2016a).

2.2.3 Machine Learning Predictions

Rather than making inferences about an unknown process, the aim of most ML work is to build a model that demonstrates a certain performance on a given task and data. As a result, automated pattern recognition techniques are used to predict unknown outcomes or future behaviour of a system based on past behaviour (IJ 2018). The fundamental assumption is the past will repeat itself. Under this premise, prediction allows for the identification of the most useful policy or action (for example, treatment selection) without needing a deep knowledge of the underlying causal processes (IJ 2018). Instead validation and monitoring of models becomes more important to avoid fatal impact of predicted decisions to an audience. This is especially true when dealing with human data, since unawareness might risk discriminating blind spots. Automatically checking models with high quality standards regarding fairness can avoid discriminatory biases, because they are hard to find and require creative and rigorous testing (more in section 2.3.1).

The goal of ML is to find a robust mapping between x values to y values. A number of parameter choices must typically be made, although learning is done inductively through automatic pattern recognition. A model is evaluated based on an empirical risk, which simply implies a dataset that the model has never seen before. Then we may assess different functions and look for the model with the lowest expected risk, that is the model with the lowest empirical risk among all models on the data. Within this process discretionary design choices underpin the logic of every model that safeguard the inductive learning process with only minimal assumptions about the data (Dangeti 2017). Inductive biases are fundamental for ML which contradicts the common notion of 'objective machines' (more in section 2.3.2).

In contrast to statistics, ML uses the process of empirical risk minimisation and sample splitting for predictive inference. During the training phase, we use a part of our data to train the model, and another part for testing. The goal is to get the

best possible performance on the test set. However, in its effort to minimize its loss function, the algorithm tends to over- or underfit to data. It might suggest overly optimistic performances by taking advantage of the noise in the data or having no predictive capability at all. Overfitting a model to a particular sample of data is also known as "reward hacking" in reinforcement learning and "p-hacking" in the social sciences, which reduces out-of-sample performance and the external validity of a model. Out of sample generalization with minimal overfitting is crucial for effective model building, yet it is commonly neglected inside statistical pipelines. Decomposing the error into bias and variance gives additional insights into how over- or underfitted a model is to the data. Overfitting for example is indicating high variance in the second component of the error decomposition. Understanding the many aspects of prediction bias helps in debugging and creating more robust low variance models (more in section 2.3.3).

Inference and prediction are different tasks, and we have to ask what best suits a specific project requirement, how to draw the appropriate conclusion from either statistics or ML, and how to make sense of empirical findings. In a research endeavor, both inference and prediction may be useful and can be combined. In the end, we want to understand how systems operate, as well as what will happen next in the future. However, the world of science does not truly progress via numerical predictions; rather, it advances through mental or formal models that allow us to infer far beyond numbers. Hence, statistical inference has been the backbone for valid conclusions from data and its underlying generating processes. Interpretability is central to statistical models, which often sacrifice performance for model insights. Their simplicity allows us to determine which factors independently contributed to a particular predicted outcome. Therefore, they are often used in the scientific process as well as in high stakes prediction scenarios like in banking or real estate. Despite better prediction results in ML, the lack of an explicit or deductive model makes it difficult to link predictions directly to theoretical knowledge about the world (IJ 2018). At the same time, predictions enable the identification of optimum courses of action in a split second (for example, treatment selection) without the need for understanding of the underlying causal processes (IJ 2018). In sum, inference and prediction are complementary in pointing us to systematically useful conclusions (IJ 2018).

2.3 A Typology of Prediction Biases

Bias in ML can have several meanings, and I will focus on three bias concepts that impact prediction quality: (1) discrimination, (2) induction, and (3) error decomposition. First, algorithmic decisions can have a detrimental impact on human lives

and thus require thorough fairness validation. As a result, this study provides several formal fairness definitions from the literature that address the vulnerability of protected groups to socioeconomic inequalities and thereby preventing discriminatory biases. Second, inductive biases are essential for a model to generalize beyond any specific dataset by learning biases inductively from data patterns. Design decisions for model families, parameters, or data transformations all contribute to inductive bias in the learning process, and we must be cognizant of our own contributions. Third, we may better understand model behavior and situations of under and overfitting by decomposing the error into bias and variance. Low variance models are preferred because their outcomes are more predictable and less impacted by random chance. Discrimination, induction, and error decomposition are fundamental concepts in ML, each representing a separate dimension and being relevant at most stages of ML development.

2.3.1 Discrimination

When dealing with human data or audiences, the idea of fairness is essential since any computation that does not adhere to this principle risks discrimination, leading to serious questions about exclusion and overgeneralization of people. Extensively validating models will not only help in reducing discriminatory biases but also simple bugs that could negatively impact the prediction quality for the whole audience. As ML systems are becoming more widespread in decision-making, the demand for fair algorithms will only increase, especially if biases happen on unlawful grounds. In the United States, for example, court rulings are often based on re-offending risk estimates generated from past arrest statistics or surveys rather than the true past and actual future social and criminal trajectory, which is mostly unknown at this point. Building models using skewed data can have considerable consequences for already discriminated subgroups, because biased data e.g. from racial policing is likely to impact pretrial risk assessment, bail out, early release, as well as sentencing and level of penalty (Angwin et al. 2016). The findings of a research released in 2016 by ProPublica clearly reveals that estimates regarding non-reoffending defenders (from a recidivism prediction tool called COMPAS created by Northpointe Inc.) varied considerably for blacks and whites (Angwin et al. 2016). Black defenders were predicted to be twice as likely to reoffend as white defenders, but they also have a twice as high rate of false positives for basic recidivism (Angwin et al. 2016). The ProPublica study prompted a public outcry regarding this flagrant violation of justice.

In practice a model may serve a great variety of people who have different feature sensitivity or misclassification rates than the model was expected to deliver. Such issues have spurred a great deal of interest in the development of fair machine-

learning algorithms. Hence, there are several fairness definitions in the literature to detect these biases and we will look at five of them. Throughout this paper, we will follow the notation of Kusner et al. 2017. Let A denote the set of protected attributes of a person, which are regardless of the definition of fairness used, assumed as a prior in every particular scenario. Furthermore, let X denote other covariates, and let Y represent the outcome to be predicted, which may be tainted by historic biases (Kusner et al. 2017). In the pretrial szenario, for example, we may set $Y = 1$ for offenders who would have committed a violent crime after being freed, and $Y = 0$ otherwise. However, y is unknown to the decision maker, who knows just the data contained in the observable attributes X and A at the time of action. Finally, $\hat{Y}$ is the prediction, a random variable that relies on A and X which is an approximation of Y through a ML model.

The first definition called fairness through unawareness can be achieved if protected attributes are not explicitly used in the prediction process (Gajane & Pechenizkiy 2017). Disparate treatment is avoided and fairness through unawareness achieved if: $P(\hat{Y}|X) = P(\hat{Y}|X, A)$ (Zafar et al. 2017). To avoid any dependency on race, gender, or age, we would simply remove any sensitive variables from the model. But even after removing sensitive variables, bias may still exist due to correlated proxy attributes, such as zip code, and some authors have called to remove said proxies from models as well (Bonchi et al. 2017; Johnson, Foster, et al. 2016; Qureshi et al. 2020). However, removing any information on protected attributes might lead to biased predictions itself, therefore protected group membership should be explicitly considered to make fair decisions (Corbett-Davies & Goel 2018). Gender-neutral risk assessments, for example, might systematically overstate recidivism risk of women, leading to overly punitive court rulings. Recognizing the predictive power of gender in this context might lead to the development of a decision rule that detains fewer persons (especially women) while providing the same public safety advantages (Corbett-Davies & Goel 2018).

The second definition - individual fairness - requires a model to generate comparable predictions for similar persons (Kusner et al. 2017). The distance measure must be properly selected, which requires knowledge of the domain beyond black-box statistical modeling. Formally, if people i and j are comparable according to a given distance metric $d(i, j)$, their predictions should be similar as well: $P(\hat{Y}|X_i, A_i) = P(\hat{Y}|X_i, A_j)$ (Kusner et al. 2017).

Third, in order to guarantee demographic parity, we have to measure how much a demographic group is impacted by indirect discrimination, and it may be detected when there is a disparity in the proportion of positive (negative) results for various sensitive groups (Kusner et al. 2017). A model fulfills demographic parity if the following condition is met: $P(\hat{Y}|A = 1) = P(\hat{Y}|A = 0)$ (Feldman et al. 2015). How-

ever, if there are real differences in base rates for two subpopulations, which might be the result of historic social and economic discrimination, it would be nonetheless inaccurate to consider an algorithm as unfair for precisely detecting existing data patterns (Corbett-Davies & Goel 2018). Furthermore, disparate impact of groups might be less meaningful in scenarios where ground truth decisions are available as for the following equality of opportunity (Zafar et al. 2017).

Fourth, equality of opportunity ensures that the probability of being in the positive class is the same for each protected group (Kusner et al. 2017). When the ground truth for previous decisions is available, it is possible to justify disproportionately beneficial results for people with specific sensitive attribute values by pointing to that ground truth (Zafar et al. 2017). That is to say, persons who are entitled to preferential treatment because of their attributes cannot be negatively discriminated against by sample averages or overgeneralization. A model is free of disparate mistreatment and fulfills equality of opportunity if: $P(\hat{Y}|A = 1, Y = 1) = P(\hat{Y}|A = 0, Y = 1)$ (Zafar et al. 2017).

Fifth, counterfactual fairness, is a technique using concepts from causal inference. The definition of counterfactual fairness captures the intuition that a decision is fair towards an individual if we get the same results (a) in the actual world and (b) in a counterfactual world where the individual belonged to another hypothetical demographic group (Kusner et al. 2017). Predictions are counterfactually fair if under any context, the prediction remains the same although there is a change in sensitive attributes from a to counterfactual $a^{'}$. This requirement can be formalized as follows: $P(\hat{Y}|X, A = a) = P(\hat{Y}|X, \text{ set } A = a^{'})$ (Kusner et al. 2017).

We have seen five definitions of fairness and how a model may be tested for discriminatory biases. Sometimes these fairness criteria are incompatible and more or less suited for certain situations (Kusner et al. 2017). For example, Chouldechova 2017 shows that various parity constraints are incompatible if base rates differ across groups. The issue of fair prediction can not be addressed without considering social issues such as unequal access to resources and social conditioning. While these factors are difficult to quantify and formalize statistically, it is important to acknowledge their impact and test predictions against various definitions of fairness (Gajane & Pechenizkiy 2017).

2.3.2 Induction

In ML inductive biases are the result of basic design choices that underpin the logic of every model. Mitchell 1980 refers to bias as a reason for preferring one model realization over another that is not based on consistency with the in-sample error. The author asserts that there is no such thing as bias-free ML and that a productive bias is required. Mitchell writes further: "Although removing all biases from a

generalization system may seem to be a desirable goal, in fact the result is nearly useless" (Mitchell 1980). The widely held belief that computers, and hence ML, are objective or free of biases is wrong, because they are a fundamental requirement of any inductive learning system. Inductive biases are assumptions and design decisions that must be fixed in order for any algorithm to derive logical conclusions from observable data. While inductive bias is an important component of any prediction model, we generally do not want noise, statistical artefacts, or discriminatory biases to be the primary drivers of model performance.

To better understand inductive biases or in general unintended biases, we can use an empirical risk minimization approach (Vapnik 1991). We can not anticipate model performance in the future since we do not know how data is generated and distributed in the future. Therefore, a model is designed to be trained on one sample and predicted on another, allowing us to approximate its unknown generalization error (real risk) with the test set error (empirical risk) (Vapnik 1999). This ensures that computational results are possibly generalized beyond a specific sample of data and will meet expectations in the future. Setting aside a hold out test dataset will also aid in the detection of vulnerabilities and arbitrary bugs before a model is considered valid and deployed. For example, take a very simple model that maps x to y and has just one free parameter. Once that single parameter has been learnt from a set of data, it is possible to evaluate the goodness of fit for a parameter. What is the generalization capacity of a model when applied to previously unknown data? What is the difference in generalization error between this model and the realistically best f^* model? To address this issue, it is common practice to divide the data into three or more parts: the train set, the validation set, and the test set (Zhang et al. 2021). On this premise, we can assess if a model that can achieve its objectives on the train set in a number of ways is still useful on a test set and does not violate said objectives. At the beginning of the splitting procedure, 20% of the data is set aside for final testing. The remaining 80% can be separated once more, this time for training (80%) and validation (20%) purposes. In this situation, splitting the data set into many folds allows us to better evaluate model performance since train and validation errors are used for optimization and selection, resulting in excessively optimistic errors compared to the true generalization error. Model selection based on the validation error may finally be back tested with the test set from the start. Otherwise, a multitude of bugs and biases could enter the ML pipeline, rendering it useless or even harmful in real-world applications. Beside measuring correct errors we will see how design choices within a ML pipeline can alter its predictive performance. To arrive at effective design choices either the factual knowledge of target domain should be sourced or methods used for automatically learning inductive biases (Baxter 2000). The initial choice of hypothesis space is

often the most difficult challenge in any ML task; it must be broad enough to include a solution to the problem at hand, while small enough to enable effective generalization from a limited number of samples (Mitchell 1980).

2.3.3 Error Decomposition

Statistical bias and variance are the two major components of error decomposition. In a nutshell, statistical bias of a learning system is the systematic error that the learning algorithm is expected to produce on average when fed training data (Dietterich & Kong 1995). Variance on the other hand measures the variability due to data sample differences or randomness in the statistical learning procedure. The error for a linear model is defined as the average difference between predicted and actual values, often referred to as the mean squared error (Dietterich & Kong 1995):

$$Error(x) = E\left[(y - \hat{f}(x))^2\right]$$

If we decompose this equation, we get

$$Error(x) = \left(E\left[\hat{f}(x)\right] - f(x)\right)^2 + E\left[(\hat{f}(x) - E[\hat{f}(x)])^2\right] + \sigma_\epsilon^2$$

$$Error(x) = Bias^2 + Variance + \text{Irreducible Error}$$

Statistical Bias quantifies the process through which a model makes simplifying assumptions in order to facilitate learning the objective function. It captures the concept of systematic error for a particular sample of data (Dietterich & Kong 1995):

$$Bias(x) = E\left[\hat{f}(x)\right] - f(x)$$

On the other hand, variance measures the amount by which the estimate would deviate from its mean if different training data were used. Formally, the variance of an algorithm is defined as the expected value of the squared difference between the prediction of each given model and their average (Dietterich & Kong 1995):

$$Variance(x) = E\left[(\hat{f}(x) - E[\hat{f}(x)])^2\right]$$

Variance describes the sensitivity of a model to variations in the training set. It is caused by random fluctuations in the data, or unpredictable behavior in the learning process itself, such as random starting weights or session seeds. The remainder of the equation is called irreducible (idiosyncratic) error as it can not be reduced regardless of what algorithm is used. Therefore, prediction error is made up of only two systematic components that may be controlled: bias and variance.

A model with a strong bias is fairly simple since the data assumptions ignore the training data and hence oversimplify the model. Since a high bias model is too simplistic but stable with respect to noise, it always results in a high error on both the train and test sets (Belkin et al. 2019). It is therefore referred to as underfitting the model to the data. On the other hand, high variance may cause an algorithm to repeat noise in the training data rather than predictive patterns. A high variance model concentrates on training data and fails to generalize to unseen data, hence it is overfitted (Belkin et al. 2019). We can always discover a model that perfectly matches all training points. However, such models tend to be highly complicated and versatile, and they also prefer to capture noisy or completely random inputs. As a result, overfitted models tend to deliver poor predictions for unknown data, with a large test error and a small train error. Overfitting and underfitting are both undesirable conditions, and a balance should be struck. As a result, one key aim in algorithm design is to minimize statistical bias and variance, and therefore the overall error. Ensemble learning strategies, especially boosting and bagging have demonstrated impressive capacities to increase the prediction accuracy by decreasing bias and especially variance (Webb & Zheng 2004). Because statistical bias and variance are influenced by the architecture of model, any modification to a ML pipeline might affect the bias and variance, as well as overfitting or underfitting.

2.4 Experiments

There are many biases that impact the quality and validity of prediction. The ubiquitous nature of prediction biases means that we need creative solutions for systematic validations. Luckily, the literature from several areas on ML validation provide us a rich tool set to deal with some of the biases. In order to demonstrate the most common biases in ML pipelines I will use real data for experimentation. The goal for this section is to build a text classifier that predicts party affiliation based on statuses from Twitter. For illustration purposes a simple baseline neural network model with only 3000 input words was used to demonstrate risks of biases in ML. The data used was collected on members of the US Congress, since for every person key attributes are known like their party registration or their gender. In the rest of the section I map out several steps and biases in a ML pipeline, as well as possible counter strategies.

2.4.1 Data Biases

Data biases are often external to, not immediately apparent or controllable by an expert. In addition to known factors, data is constantly influenced by unknown confounders unique to the conditions under which it was collected. A lack of control over

the data generation process is problematic because it raises the chance of structural data imbalances caused by measurement error, sample biases, biased collectors and annotators, or inequalities among other reasons (Zou & Schiebinger 2018). Measurement error, for example, can bias an individual's recorded outcome as well as the features that are used to predict said outcome (Corbett-Davies & Goel 2018). Bias in outcomes is perhaps the most serious obstacle facing fair ML, as censorship, survivorship and selection bias is ingrained into all human data. Moreover, algorithms should be trained on datasets that are representative of the populations to whom they will be applied. Avoiding sample bias is difficult in reality and failing to do so might have unforeseen, and perhaps discriminatory consequences (Corbett-Davies & Goel 2018).

The goal of machine-learning algorithms is to maximize overall prediction accuracy for the training data. If a certain set of people shows more often than others in the training data, the algorithm will optimize for them since it improves overall accuracy (Zou & Schiebinger 2018). Any algorithm may be given a false sense of the popularity of a majority label as a consequence of sample biases or skewed class labels. Instead, to be considered fair, the model performance must be consistent across relevant features, over- and under-representation must be appropriately regulated, and potential inequalities addressed. De-biasing the data is a crucial first step in avoiding prediction biases, and it may be achieved by balancing or removing sensitive features, misleading words from a vocabulary, or a specific harmful target class from the target variable (Zhou, Sap, et al. 2021; Bolukbasi et al. 2016). Class or sample weights, which allow for the implementation of a custom loss function, are another interesting solution to discriminatory biases (Aıvodji et al. 2019; Zhou, Kantarcioglu, et al. 2021). Depending on the cost of data, re-sampling or re-weighting the training distribution are also proposed to offset inequalities (Singh et al. 2021; Fricke & Herzberg 2019).

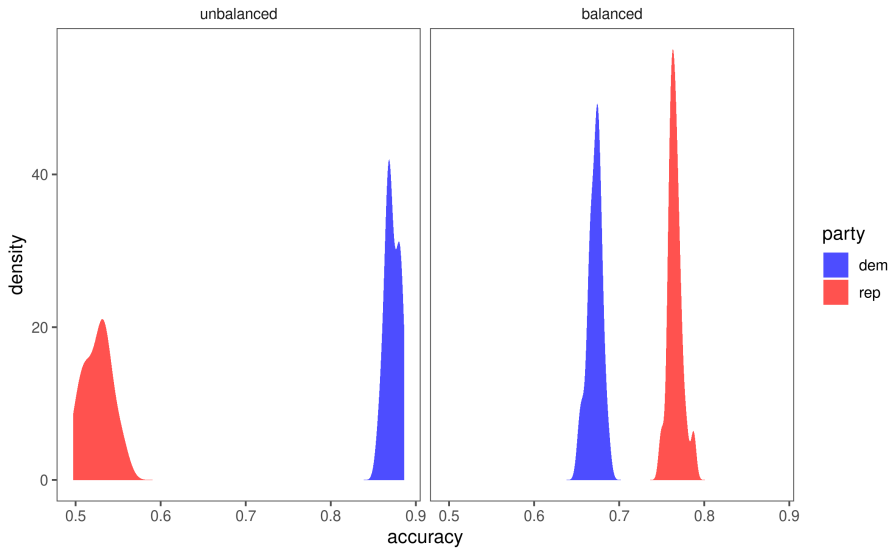
Table 1: Data Used for Experimentation Before and After Balancing

	Data	Persons	Republicans	Female	N words	N tweets
1	unbalanced	527	44%	25%	32	1387641
1	balanced	524	50%	24%	32	150000

The data used in the following ML experiments can be seen in table 1, where Republicans are slightly underrepresented (by about 6-7%) on Twitter for members of Congress (unbalanced). Furthermore, for gender we see a wide representation gap for legislators, which does neither exist in the general public nor on Twitter. The final data was down sampled from several millions to just 150.000 tweets to facilitate most computations (balanced). We can see that after balancing republicans and

democrats they are represented in the same numbers.

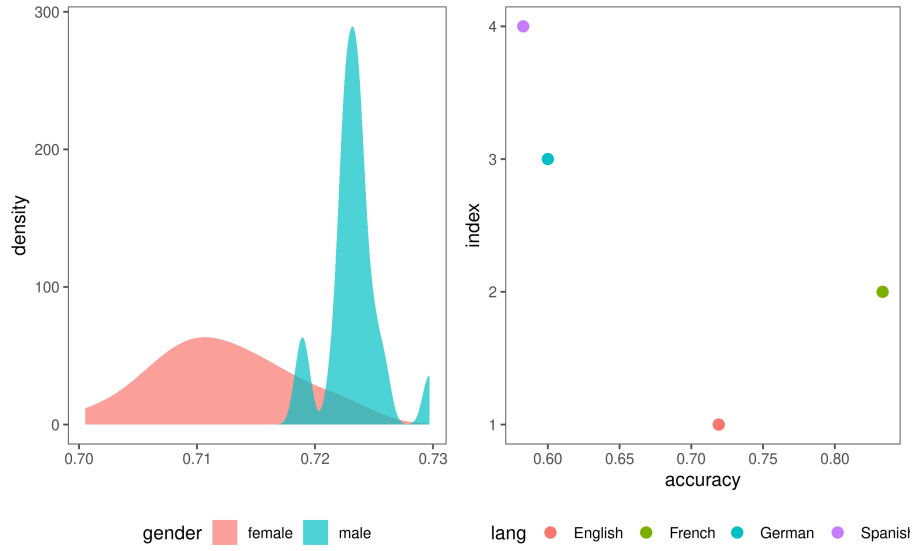
Figure 1: The Impact of Balancing Data on Baseline Model Performance



In figure 1, we can see the first rounds of bootstrapped baseline models for the prediction task. For the unbalanced data there are wide gaps in test accuracy for the target classes, raising questions on the usefulness of these models. The most serious concern is the disparity in performance between Democratic and Republican target classes, which gives the perception that Republicans are being discriminated against. Furthermore, in this circumstance, the average random benchmark accuracy is 56 %, rather than the expected 50 %, making interpretation more challenging. Balancing was undertaken for Republicans and Democrats to diminish the perception of unfairness between the target classes, resulting in a significant drop in the average performance gap between the two groups. Only now can accuracy more than 50% be readily interpreted as evidence of predictive performance. Balancing ensures that each target class contributes equally to the training process and that the demographic parity between the two parties is more accurately represented. Despite the fact that the balanced data is more useful and fair, there is still a difference between Democrats and Republicans that has tightened and reversed.

We can go even farther and look at how the models work with the balanced data for both men and women, as well as for various languages such as English, French, German, and Spanish. The gender difference in accuracy is at modest 2%, yet there is a significant representation gap for women in Congress. On the other hand, we can observe that most languages other than English underperform the most prevalent language in the sample by about 10%. This prediction quality checking has many more intriguing aspects, such as age, wealth, and other sensitive attributes. It is difficult to get such high-quality data, but in a commercial environment, you may have access to such rich information to test your models and avoid bias of all kinds.

Figure 2: Baseline Model Performance for different Sub-groups



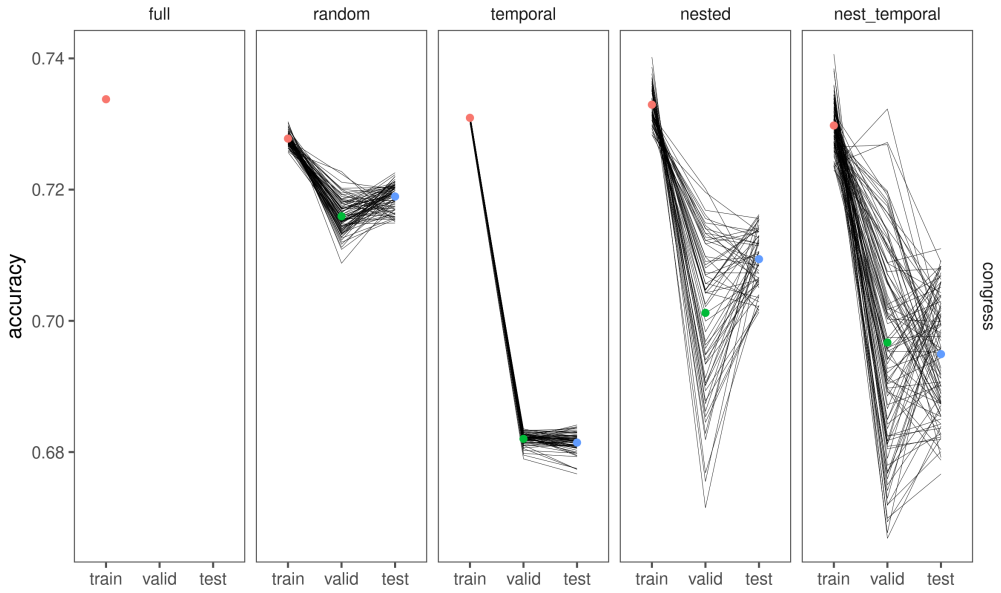
We may be tempted to minimize prediction bias by post-processing the trained model, by adding a calibration layer to the output of a model. By addressing the symptom rather than the fundamental problem, we build more complex and unstable systems that must be maintained. Projects that use calibration layers tend to become too dependent on them, relying on calibration layers to correct all model flaws.

2.4.2 Evaluation Biases

To avoid the danger of overfitting or underfitting a model to a particular sample of data, inductive design choices must be subjected to systematic testing. ML models are typically evaluated by the predictive power and validity of a model using empirical risk minimisation and data splitting into multiple folds for out of sample testing (Vapnik 1991). Four splitting techniques will be demonstrated, that allow to quantify bias and variance within an evaluation process: (1) randomly splitting, (2) temporal splitting along the time axis, (3) nested splitting alongside repeated units of observation, and (4) nested-temporal splitting accounts for time and repeated units of observation. These four splitting techniques will be bootstrapped to provide an inherent perspective on sample uncertainty and thus model bias and variance. In the following experiments, all model-related parameters are maintained constant to illustrate simply the sample uncertainty arising from correctly splitting and evaluating data and models. The baseline model received the top 3000 words from the training corpus (N = 150.000).

In figure 3 we can see that the bias (train error) is fairly stable over all experiments. But as we introduce sample splitting the empirical risk increases and

Figure 3: The Impact of Sample Splitting on Baseline Performance

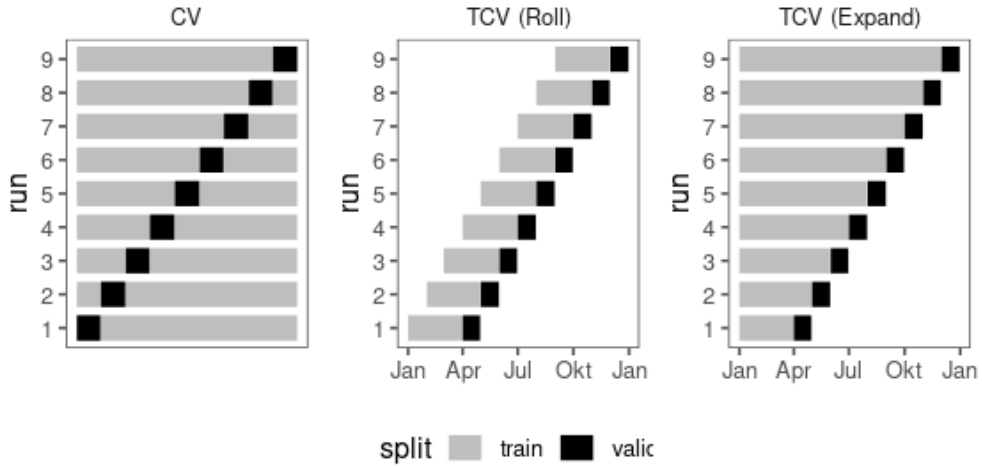


accuracy decreases. Keeping track of the train error alone may be challenging, as we can observe from all trials that we get too optimistic performance estimates. Randomly splitting the data might help to correct initial expectations, but it may conceal more complicated patterns in the data. When we examine nested observations, we add more variance across the experiments since the unit of analysis are the tweets of a person, which are more likely to have lower within and greater between variation. Furthermore, splitting the data by time reveals that the model performance suffers even more since recent observations cannot be used to predict older ones, preventing information leaking between time frames and increasing variance. When nested and temporal splitting are utilized simultaneously, most spillovers or leakages are avoided, resulting in the highest possible performance correction in the majority of trials. Overall, these splitting techniques aid in model evaluation, comparison and selection, as well as allow us to adjust our expectations for future out-of-sample performance.

A more comprehensive method called cross validation is a well-known advanced statistical resampling technique for evaluating the risk of a given statistical model and selecting from a family of models (Mervyn Stone 1974; Stone 1978; Arlot & Lerasle 2016). Two well-known approaches will be discussed in this section: k-fold (CV) and time series cross validation (TCV) (Bergmeir & Benıtez 2012). CV divides the data into k folds, as shown in figure 5, and repeatedly uses one fold for validation and the remaining k-1 folds for training. As a result, the cross validation error may be represented as the average validation error over all folds, as follows (Arlot & Lerasle 2016):

$$CV(f) = \frac{1}{N} \sum_{k=1}^{K} E(f_k)$$

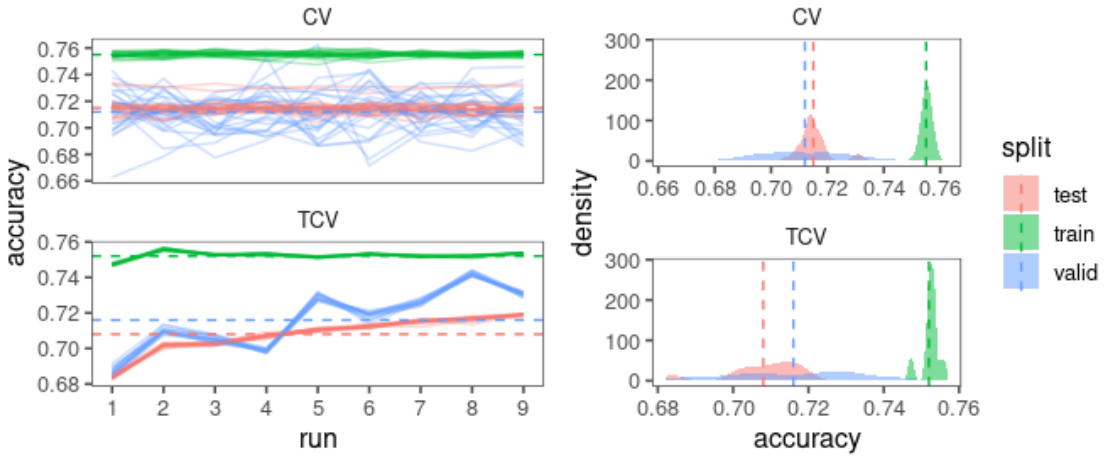
Figure 4: Cross-Validation Schemes for Predictive Models



When compared to simply dividing the data into train, valid, and test parts in figure 4, this approach makes greater use of the data because each data point contributes to the error computation at some point. k is an additional hyperparameter that must be selected before or as part of model tuning. The number of folds k should be set between 5 and 10 because more folds do not necessarily reduce variance or have an impact on model risk assessment (Arlot & Lerasle 2016). A smaller dataset would require leave-one-out validation, in which a model is fit repeatedly to the whole data set with the exception of one data point for validation. As a result, we would have the same number of models as observations in the dataset, which is not acceptable for the size of the data in this project. Despite the fact that the K-fold CV is computationally expensive (especially during model tuning) and still underestimates the true risk or generalization error, there is no better systematic validation procedure than some variant of CV. The following figure 5 shows a repeated nested k-fold cross validation with 9 folds for the baseline model on both datasets. That means in every fold there are only tweets from a selected number of people not overlaying to other folds. Nesting ensures basically that repeated observations from persons are block sampled and not leaked between train, valid and test set. The bootstrapped boundaries yield [68%; 74%] with an average performance of 72% accuracy.

When the data distributions are neither stable nor independent, a more complex split validation technique is required (Zhang et al. 2021). For example, a model could learn temporal correlations between topics being discussed and the news cycles, which would violate the assumption of independence in several ways. Time

Figure 5: k-Fold Cross-Validation with a Text Model



Series Cross Validation is a step up in sophistication since it accurately accounts for changes in the underlying data distribution and allows for the generation of dynamic systems that are biased toward more recent data. Depending on the setup, any model that has been trained to account for new data must be retrained. TCV must be introduced to a modeling pipeline with at least three additional parameters in order to be effective: the train and valid window, as well as the amount of data to incorporate in the first place. Another factor to consider is whether the training window should be expanded or rolled forward with newer observations as time passes. TCV is very useful when dealing with non-stationary data challenges, such as time series prediction or data from the newscycle on social media, where regime/concept drifts must be considered and leakage of information between time frames must be prevented at all costs. Using new data to predict old ones just inflates the evaluation metrics upwards with no merits in the future. To avoid leaking future data into the past, the model is retrained on a monthly basis up to t in order to predict $t + 1$ rolling forward. TCV gives the most realistic estimation of the generalization capabilities of a model. In figure 5 can identify for TCV draw downs and fluctuations in monthly performances of up to 5% accuracy and a steady positive trend over time.

In general, the more parameter a model has or the more non-stationary a data problem is, the more likely we want to use sophisticated resampling techniques in order to arrive at realistic model evaluation that leads to robust model selection. The main objective should never be to select a model with the highest accuracy or reward, but rather weighted by the risk of failure, hence improving outcome stability. By using monthly evaluations we avoid selecting a model that performs extraordinarily well in one month but worse than average on the rest of the months. We rather should select a model with a high monthly average risk adjusted reward

ratio to better guarantee out of sample performance going forward.

2.4.3 Transformation Biases

Data preparations are usually transformations of the training set that may involve substantial data cleaning and filtering operations. This is often the first stage in a ML pipeline, and it can make or break the predictive abilities of a model (Kuhn & Johnson 2013). Different models are more or less sensitive to the predictors in the model and data transformations may dramatically enhance performance by reducing the reliance on and effect of data skewness, outliers, and missing values. It is also known as feature engineering since it allows you to enrich a given dataset with better data representations of the underlying data. Consider a function that just takes x and maps to the outcome y like $y = f(x)$. If we use a constant log transformation to normalize the data, the model becomes $y = f(log(x))$. It is simple to create a structurally different model without introducing any additional free parameters. Minor changes to data or model, can have a significant impact on performance and model logic. The decision to include the constant transformation is a design decision or macro parameter in and of itself that must be monitored and, if necessary, tuned. If we use a min max normalization, the data is normalized into [0; 1] ranges and may be better demonstrated as a separate model:

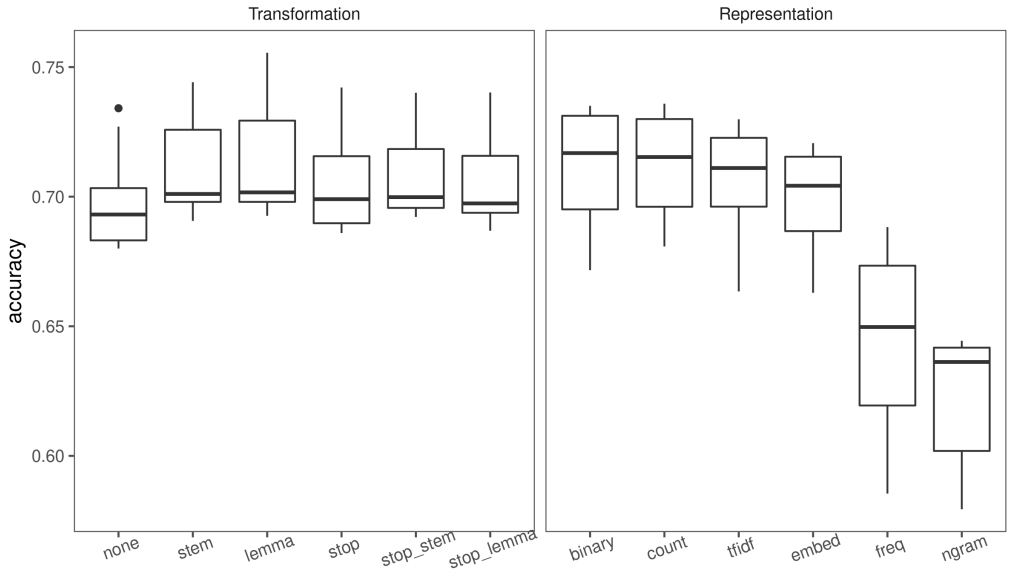
$$\delta(x) = \frac{x - min(x)}{max(x) - min(x)}$$
$$y = f(\delta(x))$$

As a consequence, the final model incorporates sample estimates of min_{train} and max_{train} from the underlying data. They must be regarded as direct model parameters that add to model complexity. Because model performance is dependent on previous sample estimates, they must be treated as if they were part of the model itself. To address this transformation issue, every optimisation pipeline should be structured to work like a cooking recipe, with each transformation performed prior to fitting being recorded. This is implemented in R by the package "recipes" (Kuhn & Wickham 2020) or sklearn pipelines (Feurer et al. 2019) in python. In the case of tabular data, this might include coping with missing values, which can lead to unpredictable behavior or even the entire collapse of a model. In order to impute future missing data, we can use sample estimates from the training data, such as the mean or median. This improves the model stability and robustness, but it also necessitates rigorous testing and validation. The logic of model stacking also applies to imputation in the predictive domain. Each successive model layer brings new assumptions and parameters into the modeling pipeline, which must be validated outside of the training sample to confirm they are useful in comparison to simpler

techniques.

Any kind of data transformation or parameter choice can increase a model complexity. Especially estimates like scaling or normalizing the data based on sample estimates like min and max have to be treated as part of the model. This means for text classification that the entire tokenization process is part of the model itself that can be tuned. Although generally we do not have to deal with missing data (due to a priory filter for nwords > 5) there are several parameters learned from the text of the training data. This includes the top N words in the vocabulary as well as sample estimates for term and document frequencies (TFIDF), or word normalizations like stemming or lemmatization.

Figure 6: Text Transformations and Representations and their Impact on Test Performance



The following figure 6 shows the test performance for various text transformations and the baseline neural network with 3000 words in the vocabulary. Although rather small, any kind of word normalization is outperforming the scenario in which no transformation was applied. Particularly useful appears to be lemmatization which replaces a word with its word root, and therefore reduces the overall size of the vocabulary as more words can be represented in the same input space. In contrast, removing static lists of stopwords can be counterproductive and results are mixed. After cleaning the texts they have to be shaped to some computer readable format which is traditionally a document term matrix. In figure 6 on the right hand side, we can see different text representations and their average test performance. It is quite obvious that ngrams (two word occurrences at once) is failing to perform in comparison to the standard binary or count approach. Moreover, trained word

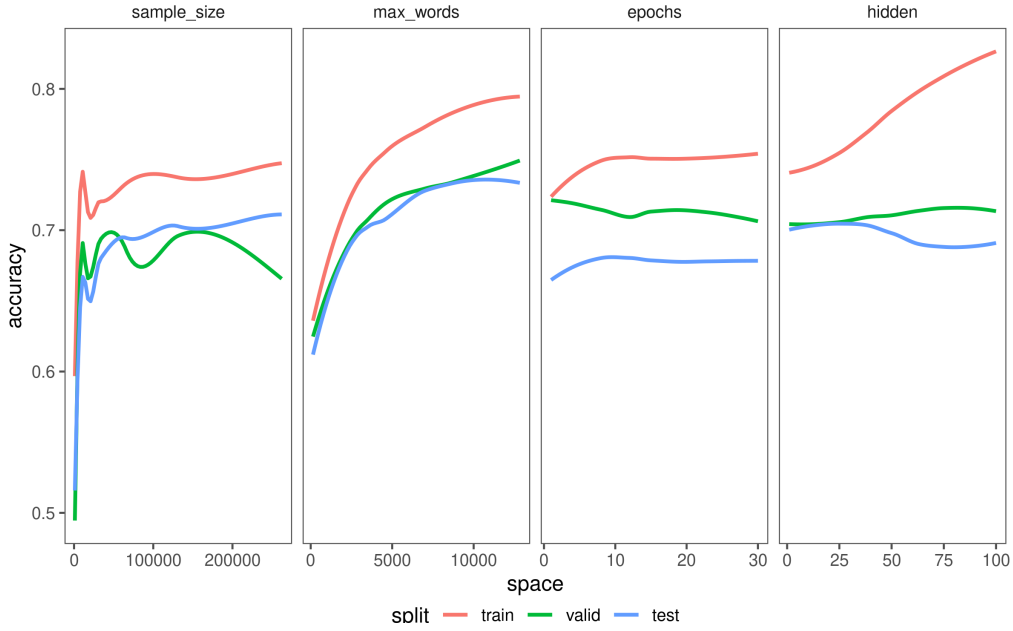
embeddings (not pretrained) with the baseline classifier is performing slightly worse than the standard approaches of just counting the words for each document.

2.4.4 Model Biases

Any bias introduced into a model through data, sampling, or transformation has the potential to be preserved, if not exacerbated. The interplay of data and model design decisions is the most common cause of bias in a system (Hooker 2021). It is feasible to avoid bias in this circumstance if the impact of the model design is understood. This is due to the fact that algorithms are not neutral, and certain design choices are more preferable to others (Hooker 2021). The ultimate goal is to build a prediction model with low bias and low variance which are strongly impacted by data and model complexity.

Data complexity refers to the number of training instances supplied. Even if your model is simple, a dataset with only one or two examples is quite easy to overfit. Overfitting a dataset with millions of samples, on the other hand, requires the use of a highly flexible model (Ying 2019). As we increase the amount of training data, the generalization error typically decreases and more high quality data never hurt. For a fixed task and data distribution, there is typically a relationship between model complexity and dataset size. Given more data, we might attempt to fit a more complex model.

Figure 7: Impact of Parameter Choices on Baseline Model Performance

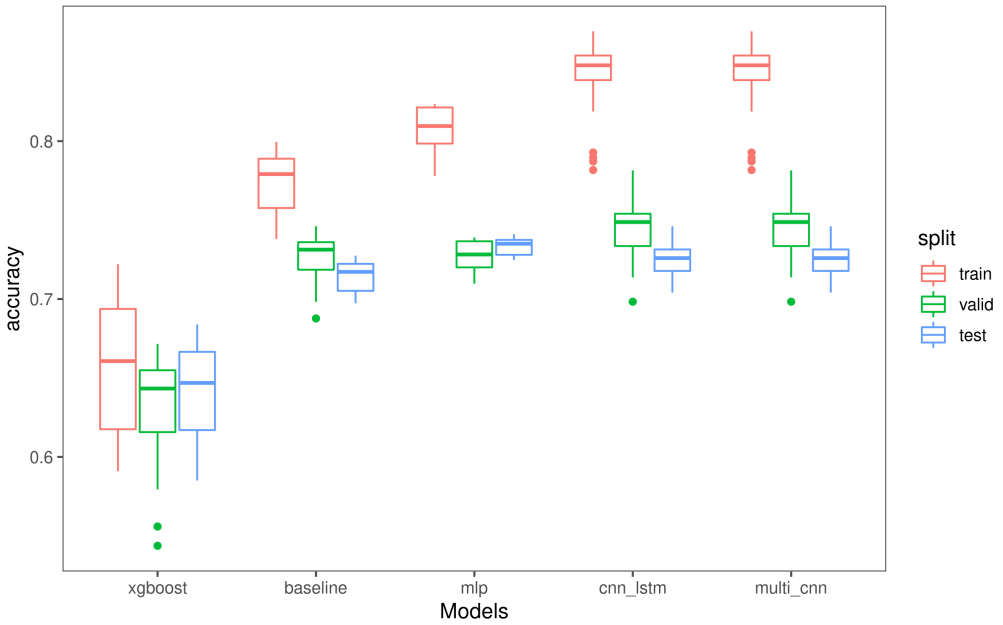


According to experiments shown in figure 7 for sample_size, a few thousand tweets are required to establish performance stability for the baseline model, with a higher number of tweets recommended. Second, the more data points analyzed, the

closer the performance metrics from the train, valid, and test sets converge. Based on a document term matrix with 3000 features (words) in the input matrix, a 64GB 32core server cannot handle more than 300.000 tweets at the same time, resulting in resource exhaustion and a stack overflow. This is one of the primary disadvantages of the classic document term matrix approach, which has been rectified by the development of more complex neural networks, which rely on much denser input data and can thus handle millions of samples in seconds, as compared to the traditional tabular ML approach.

Another important factor for performance is how many features (words) are known to the model. From figure 7 max_words, we can see that the more words are free to learn the better the performance gets with a clear saturation effect for this particular baseline model. Above 15000 words and 150000 instances to train on the machine fails to compute predictions at once. The number of estimated model parameters is referred to as model complexity. Models grow increasingly prone to overfitting as the number of configurable parameters rises, which is undesirable. Not only is the quantity of estimated parameters important, but so is their scaler range. Whereas the number of epochs a model is fed the data again and again does not seem to cause prediction instability, the number of hidden units beyond 50 is doing so. While the training accuracy increases with more hidden units, the empirical risk of valid and test sets remain the same or decay. That means we overfit the model to the data with increased hidden units without merits in future deployment.

Figure 8: Impact of Model Family on Performance



Lastly, as seen in figure 8 model bias and variance can be influenced by the choice of the model family. From interpretable statistical models to giant neural

networks any learning algorithm could be used to predict this task, although with varying efficiency. For the following experiments I settled on five model families namely (1) xgboost: a boosted tree ensemble, (2) the baseline neural network with one layer, (3) mlp: a multi layer perceptron with relu activation, (4) a stack of CNN and LSTM layer and (5) a multi channel CNN architecture. Xgboost is a powerful tree ensemble that wins almost any tabular data challenge. However on sparse wide input data as it is the case for text input the model is struggling to perform well. Even the baseline model is doing way better than xgboost. Mlp is not outperforming the baseline model significantly which means only a minor increase in test accuracy. The two most complex models with several thousand parameters ccnlstm and multicnn are designed for use cases with much more data than in these experiments for comparison purposes. They perform equally on medium sized data but eventually outperform on vast quantities of data, the statistical models can not even process. Compared to XGBoost all neural networks have way more parameters and are more prone to overfitting but they also show off higher test performance.

2.4.5 Execution Biases

Execution biases occur if there is a difference in training or evaluation environment and the production environment. This could be due to shifts in features, concept drifts over time or code bugs that change the conditions a model is used for prediction. This can be avoided through live monitoring and post-authorization of model updates over time (Vokinger et al. 2021).

2.5 Conclusion

The internet, big data, and statistics have had a profound impact on many disciplines of study and aspects of life, and machine learning is no exception. Machine learning offers enormous potential in the social sciences, such as massive data discovery, measurement, and treatment prediction. However, before transferring any responsibility to a decision-making algorithm, regardless of the context of an ML solution, a number of security risks must be addressed, the majority of which are related to bias. As ML is more and more used by non ML experts, this article tries to provide guidance, as well as strategies for dealing with the most common biases in practice. As shown by the literature, the simple goodness of fit is not enough to judge the predictive quality and fairness of a specific model. In this paper, I propose a new typology of biases and distinguish between between discrimination, induction and error decomposition. Finally, I showed how those types of biases matter at each step of a classical ML pipeline and mention possible mitigation strategies. As such, it is critical to reevaluate past research on biases in ML and conduct experiments,

to identify best practices moving forward. The main objective of this paper is educational and should be seen as a practical guide to bias in machine learning. After all, a more comprehensive understanding of what contributes to biased predictions is important because it determines where we focus our efforts to mitigate potential harm caused by biased predictions and how robust machine learning models are built in the first place. Machine Learning is transforming economies and societies with long endured inequalities and our models should not sustain or worsen it (Zou & Schiebinger 2018).

3 Personalization Bias and Twitter: How do content recommendations respond to ideological behavior?

Benjamin Guinaudeau, Simon Roth and Fabio Votta

Abstract

Although social media only recently emerged, the accumulation of evidence undermining the 'echo chamber' hypothesis is striking. While self-selective exposure to congruent content - the echo chamber - is not as salient as expected, the ideological bias induced primarily by algorithmic selection - the filter bubble - has been less scrutinized in the literature. In this study, we propose a new experimental research design to investigate recommender systems. To avoid any behavioral confounder, we rely on automated agents, which 'treat' the algorithm with ideological and behavioral cues. For each agent, we compare the ideological slant of the recommended timeline with the ideological slant of an artificially reconstructed chronological timeline and, hence, isolate the ideological bias of the recommender system. This allows us to investigate two main questions : (1) how much bias is induced by the recommender system? (2) what role is played by implicit and explicit cues, when triggering ideological recommendations?

The pre-registered experiment [1] features 170 automated agents, which were active for three weeks before and three weeks after the 2020 American presidential election. We find that, after three weeks of delivering ideological cues (following accounts and interacting with the content), the average algorithmic bias is about 5%. In other words, the timeline as structured by the algorithm entails 5% less cross-cutting content than it does when it is structured chronologically. While the algorithm relies on both implicit and explicit cues to formulate recommendations, the effect of implicit cues is significantly stronger. This study is, up to our knowledge, the first experimental assessment of the ideological bias induced by the recommender system of a major social media platform. Recommendations rely above all on behavioral cues unwarily and passively shared by the user. As affective polarization becomes a greater contemporary challenge, our results raise important normative questions about the possibility of opting-out from the ideological bias of recommender systems. In addition, it points out that more transparency is urgently needed around the recommendation questions: How are algorithms trained? What cues or features do they use? Against which biases have they been tested? In parallel, the results demonstrate the failure of 'in-house bias correction' and calls for an external auditing framework, that would facilitate this kind of research and crowd-source the scrutiny of recommender systems.

[1] https://osf.io/5kwpr

3.1 Introduction

Social media and the role it plays in our lives has evolved drastically in the past decade. With the advent of YouTube, Twitter and Instagram, the social media world has further diversified since it's inception of its modern variant with the launch of Facebook in 2004. All grown up now in 2020, the new social media hit TikTok is one of the most downloaded smartphone applications in the world. The numbers are impressive: an average American spends more than 2 hours a day on social media (Statista 2019). Even though each platform has its own affordances, they all need to solve the problem of content oversupply. Social media differs from classical media outlets, such as newspapers or television, because it democratizes content production. Paired with increasing access to the internet and availability of smartphones, social media has dramatically sunk the cost of producing online content in the last twenty years. Almost anybody can now start a YouTube channel, comment on politics on Twitter or publish pictures on Instagram. In 2015, Youtube estimated that 500 hours of videos were uploaded every minute. A year before Twitter declared that about 500 million tweets were produced every day. These mind-boggling numbers go far beyond what any human could consume, even if they would spend 100% of their time on social media.

To facilitate navigation of the enormous volume of content, each social media platform had to adopt a specific set of tools. Each platform features some sort of opt-in mechanism, where users can select specific content or content producers (for example, by becoming "friends" on Facebook, "following" on Twitter or on Instagram, "subscribing" on YouTube or Twitch). However, as platforms incentivize producers to create a lot of content, the opted-in volume can still exceed any reasonable consumption capacity. Therefore, on top of the active selection performed by the users, social media platforms use algorithmic recommendation to reduce the content supply to a manageable size.

Recommender systems (RS) deployed by social media, follow one main purpose: optimizing user experience on the platform. The opacity around these tools makes their description very hard. What model do they rely on? What kind of data is the RS trained on? Which metrics are used for optimization? To what extent are they tested for inherent bias and what is implemented to mitigate such biases? Most of the academic literature dealing with RS aims at improving the recommendation and very few contributions deals empirically with testing actual and deployed RS. In a Twitter blog post from 2020, two machine learning engineers at Twitter acknowledge that RS "aim to maximize user satisfaction as well as other key business objectives" (Twitter 2020b). Another blog post from 2019 indicates that user engagement is one of the core metrics on which RS are trained (Twitter 2019b). In 2019, Ashish Bansal,

Senior Engineer Manager in charge of recommendations at Twitter, gave a talk on challenges of scaling up recommender systems (Ashish Bansal 2019). He states that collaborative filtering is particularly well-suited for Twitter data (in opposition to content-based recommendation). Finally, a blog post from 2017 presents in greater detail how Twitter recommends tweets (Twitter 2017). We learn here that models are informed by "a list of considered features and their varied interactions", which includes the tweet itself, its author and the consumer. Once again, the post speaks of a "set of metrics we use here usually relate more directly to usage and enjoyment of Twitter". To summarize, Twitter only discloses very imprecise and vague information on its recommender system. In the absence of more details, we must rely on guessing to formulate expectations related to the behavior of Twitter's RS.

Because a recommender system sorts and prioritizes content, it can introduce bias and systematically favour or penalize specific content. For example in September 2020, the picture cropping system used by Twitter was the subject of a controversy. This algorithm was meant to automatically segment uploaded pictures and produce a standardized preview that would fit well with Twitter's user interface. When testing the cropping system with pictures representing the portraits of several persons, the algorithm proved to systematically emphasize white people over people of color. In a response to the controversy, Twitter's Chief Design Officer Dantley Davis, explained that they tested the algorithm for such bias. Right after he claims: "while our analyses to date haven't shown racial or gender bias, we recognize that the way we automatically crop photos means there is a potential for harm" (Twitter 2020a). This anecdote shows how discriminatory recomendation bias can emerge, even when there is no intent of discrimination. Recommender systems are trained on past human behavior or digital traces and therefore reproduce human bias. This example raises many normative questions about the potential bias of recommender systems: Beyond racial bias, what kind of other bias could algorithmic recommendations harbor? How can we properly test for bias? And if biases are found, how can social media platforms be held accountable for their tools? To what extent do RS not only reproduce but also amplify bias?

Because they are trained on data, algorithms are seen as less biased and more "rational" than humans (van der Voort et al. 2019). In reality, they appear to reproduce human biases and should be investigated accordingly. This is especially the case because the induced bias has the potential of influencing peoples' behavior and perception. As such, quantifying recommendation bias is a first step towards understanding how recommendation bias emerges and the role it plays in our lives. The present study focuses on Twitter and ideology and proposes an innovative experimental research design to measure recommendation bias. In doing so, we seek to answer two research questions: to what extent do Twitter's recommendations

respond to stated ideological preferences? What types of cues are used by Twitter to feed their recommender system? The paper contributes to the literature on social media in three different ways. First, the experimental setting relies on automated agents instead of humans to strictly control the online behavior. In doing so, it is much easier to isolate recommendation bias from confounding human behaviors. Second, we show that 'filter bubbles' exist. Three weeks of ideologically-sided interaction reduced the proportion of recommended crosscutting content by 5%. Interactions such as liking or retweeting matters more for recommendations than simply following other accounts. Finally and as discussed in our conclusion, we want to highlight the transparency of social media platforms in regard to their recommendation system and call for the development of a dedicated framework allowing independent researchers to properly audit the algorithms. The rest of this paper proceeds as follows. After discussing the relevance of investigating ideological bias, we draw two conceptual distinctions: (1) self-selective exposure vs. algorithmic personalization and (2) implicit vs. explicit cues. We then turn to our experimental design involving automated agents. The experiment was pre-registered, before its implementation during the 2020 presidential election. Finally, we present and discuss ours results.

3.2 Theory

The ideological homogeneity of online spaces as well as its potential influence on online and offline political behaviors have already been broadly discussed in the literature. Very often, it is related to the concepts of "echo chambers" and "filter bubbles". These two concepts are similar in their consequences, but must be kept apart because they describe very different mechanisms.

3.2.1 Echo Chambers

Scholars use the concept of "echo chambers" to characterize situations in which individuals only receive "echoes of their own voices" (C. R. Sunstein 2017). This can happen if a person only frequents groups of friends with similar interests or always consumes the same sources for news and information.

Researchers have, in the last years heavily investigated, how echo chambers influence the attitudes of individuals and their political behavior. Studies conducted by (Tewksbury 2003, Garrett 2009, Beam 2014) confirm the assumption that users tend to access content that underpins their political attitudes. In a study that measures the effects of exposure to congruent partisan content in the US and Israel, evidence suggests that exposure to supportive information increases affective polarization (Garrett, Gvirsman, et al. 2014) and polarization via homogeneous per-

sonal networks (Druckman et al. 2018). Others have found that the selective use of congruent media does not necessarily lead to the active avoidance of incongruous content (Garrett, Carnahan, et al. 2013).

A review by Zuiderveen et al. (2016) investigates the current empirical evidence of echo chamber effects on political opinions (Zuiderveen Borgesius et al. 2016). While it appears that there are measurable and statistically significant effects of echo chambers on political attitudes of individuals, they remain rather low or moderate. The authors conclude that there is no empirical evidence that would warrant great concern about echo chambers. One of the more recent systematic assessments of the filter bubble hypothesis comes from Möller et al. (2018), who test content diversity among news stories selected by human editors and multiple set-ups of automatized recommendation systems (Möller et al. 2018). The authors conclude that both human and algorithmic methods of prioritization provided a good diversity of opinion. There is also debate in the literature on whether echo chambers are a problem at all. For example, Guess et al. 2018 describe the somewhat limited evidence of echo chambers online and even talk about an "echo chamber about echo chambers" where people lament the supposed problematic nature of echo chambers without considering systematic evidence in the academic literature (A. Guess et al. 2018).

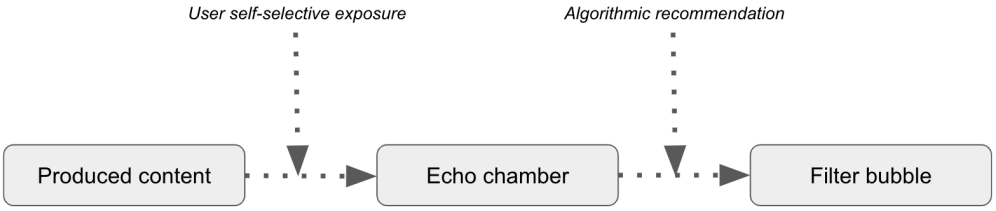
3.2.2 Filter bubbles

Often conflated with the concept of "echo chamber", filter bubbles actually describe a different mechanism. Echo chambers explains the emergence of homogeneous political environments by individual's tendency to expose themselves to content, reinforcing, or confirming their views and avoiding sources that challenge them (Beam 2014, Frimer et al. 2017). Filter bubbles, on the contrary, are not caused by self-selective exposure but are the result of algorithmic personalization. If recommending algorithms rank content, so that it pleases the user, this also reduces the likelihood of crosscutting content appearing in the timeline. In the end, echo chambers and filter bubbles produce very similar outcomes: they control what kind of content a specific user is exposed to. The main difference refers to the authority in charge of selection. While in an echo chamber, users actively select the content - or the producers - they are exposed to, in the case of filter bubbles, users do not directly control the selection process, which is taken over by an algorithm. Users benefit however from indirect control over the selection, as their behavior influences algorithmic recommendation. One mechanism does not preclude the other and they can co-exist without any problem. A user can pre-select an ideologically diverse pool of content producers, but still end up in a homogeneous space because of algorithmic personalization. The assumption that these two mechanisms act independently from

each other is the foremost motivation of this study, whose design is meant to isolate filter bubbles from echo chambers.

We know painfully little on the systems that drive algorithmic recommendation on social media. As they are meant to personalize content, we assume they rely on personal information shared by users. Each behavior on the platform, but also outside the platform thanks to cookies, can thus constitute an information cue feeding the algorithm. The algorithm is certainly trained to maximize one or several metrics, such as the time spent on the platform, the engagement of users or the generated advertising revenue. This metric is very relevant and can dramatically affect the types of recommendation. An algorithm trained to maximize engagement could for instance foster counter-attitudinal content to generate conflict and mobilize users. This would come at the risk of reducing the pleasure of users, who may eventually reduce the time spent on the platform.

Figure 9: Conceptualization of exposure mechanisms



Focusing on ideology, many different cues can reveal policy preferences, as individuals adopt behavior that are typical of specific ideological positions. For instance, they follow and interact in a gentle manner with other individuals that share their views. Conversely, they are likely to adopt a harsh tone with individuals holding counter-attitudinal views. When they click, pause, wait, open a link, watch a video for the second time, follow or unfollow accounts, they provide informative cues about their preferences. Taken all together, the complete set of online micro-behaviors we adopt online provides a precise picture about our preferences, which can be exploited by algorithms to perform recommendations. We expect the recommender system to maximize the exposure of users to content they agree with - denoted as congruent in the rest of the paper.

3.2.3 Implicit vs. Explicit Cues

Now, because of the technical limitations (structural complexity, computational costs of training and deployment) surrounding the development of recommender systems, each conceivable behavioral cues is unlikely to be considered by the algorithm. Platforms sometimes explicitly ask their users about their preferences or interests. In a sense, when users opt-in and select content producers, they explicitly

state their preference towards this content producer. Directly asking users to opt-in for specific content is certainly the easiest way to obtain their preferences, but this is also constrained as users spend most of their time on a platform consuming content and not answering questions or opting-in for more content. To improve their personalization, platforms have hence incentives to include other types of behavioral cues, such as clicks, mouse movements or browsing data. We denote these cues as implicit, because the corresponding behavior does not aim explicitly at selecting content (in contrast to answering questions about interest and following certain accounts). Users, when browsing social media platforms, continuously and unawarely generate implicit cues. As such, these cues are much more numerous than explicit ones but also much less informative. For instance, if a user explicitly mentions an interest for the republican party or follows many republican members of congress, his conservative leaning is straightforward. In contrast, if the same user scrolls faster on a democrat tweet or visits Fox news once, not much can be concluded about his leaning as democratic-leaning users might also behave that way. However, as this user keeps going back on Fox news and systematically skips each additional democrat tweet, his ideological leaning becomes clearer.

Cues do not differ only in their technicality, but also in their underlying intention. As mentioned, explicit cues aim at selecting content, while implicit cues are generated without the purpose of selecting content. This has normative consequences as unaware personalization can bias how a user perceives its environment. Elia Powers collected a small sample size of 147 college students, which revealed that they were mostly unaware of personalization techniques on the platforms they use (Powers 2017). Other (small scale) studies find similar results when individuals are asked how their (social media) news feeds are organized (Eslami et al. 2015, Rader & Gray 2015). This unawareness could lead users to think they obtain a representative view of what is produced on a platform, and, by extension, the public forum. For instance, an pro-life activist following people from both sides, may because of implicit personalization only encounter pro-life content. He may then underestimate the proportion of pro-choice users or perceive pro-choice voices as silent. Although we do not expect implicit cues to trigger more personalization than explicit ones, the underlying unawareness could have bad normative consequences. Accordingly, this study does not only attempt to measure personalization bias, but also to attribute this bias to implicit and explicit cues.

3.3 Methods

In order to explore the amount of bias as well as the respective roles played by implicit and explicit cues, we design an experiment, where we use automated Twit-

ter agents to treat Twitter's algorithm with implicit and explicit cues. We chose to focus on Twitter, which is one of the leading social-platform for micro-blogging (TwitterGov 2016, Twitter 2019a). Additionally, Twitter's data API made it reasonably easy to access the full tweet history of the content producer encountered in our sample, which eased the synthetic reconstruction of chronological timelines as described later. Lastly, the existence of short keys to navigate Twitter facilitated the automation of agents.

3.3.1 Infrastructure

To investigate Twitter's algorithm, we generated 170 Twitter accounts [2] that were randomly allocated to one of the eight treatment groups listed below. Each agent consisted of a random name, a random avatar image [3], and a real cell phone number. To prevent a geographical spill-over between agents, we also allocate a VPN location to each agent to ensure that two separate agents do not have the same IP when connecting to Twitter. Figure 10 presents a screenshot of an agent, accompanied by the metadata that we randomly assigned. [4] During all the experiment and up to six times a day, each agent will (1) log in, (2) scroll down for about 50 tweets and, if necessary, (3) follow the twitter account assigned to it.

Figure 10: Example Account and Meta Data


```
[
  {
    "user": "christman34",
    "firstname": "Yuki",
    "middlename": "",
    "lastname": "Christman",
    "password": "...",
    "sim_id": 198,
    "phone": "+447459842494",
    "mail": "yuki.christman.2503",
    "liberal": true,
    "explicit": true,
    "implicit": true,
    "rpi_index": 2,
    "tz": "America/Los_Angeles",
    "country": "United States",
    "city": "San Jose",
    "region": "California",
    "vpn": "us5304.nordvpn.com.tcp443.ovpn",
    "host": "192.168.1.15",
    "profile": "ChristmanYuki"
  }
]
```


The experiment is separated into two stages, which took three weeks each. The first lasted from October, 13th 2020 to November, 3rd 2020 and was only used to

[2] Although pre-registration mentioned 300 agents, we only managed to create just over 200 agents. Several technical reasons caused the attrition of 30 agents: banned by Twitter, loss password, malfunctioning sim card, etc...

[3] Pictures were randomly drawn from the website 'thispersondoesnotexist.com'

[4] To prevent Twitter from linking the agents together, we also randomized connection time and the user agents of the browsers used to connect to Twitter.

administer treatments (treatment phase). The second phase began on November, 3rd and finished on November, 24th 2020 (measurement phase). During this period, the treatments stopped (no additional following and no interaction) and we simply collected the recommended timeline without any further manipulation. The results presented in the next section correspond to the data collected in this second phase. We also uniquely focused on the American political context, which was thriving with political content during the experiment because of the 2020 presidential election. The bipartisanship of American politics made the measure of ideology as well as the assignment of ideological treatments more manageable. Since American politics is mostly organized around the conservative-liberal opposition, we assign one ideological side to each agent and deem all content from the other side to be counter-attitudinal. Thus, immediately after being created, each agent was first given a random ideological position: 'Liberal' or 'Conservative.'

To administer implicit and explicit cues, we took advantage of the structure of Twitter. Twitter main feed consists of the content produced, shared or liked by a pool of accounts exclusively selected by the user (followed accounts). New users have first to actively select ("follow") accounts. Eventually, the content generated by the chosen accounts will end up on the timeline of these new users. The action of following an account is seen as an explicit cue, because it actively and explicitly selects a content producer. Agents assigned to the explicit treatment would provide explicit cues about their ideology by following 80% of congruent accounts and 20% of cross-cutting accounts. So for example, a conservative agent would follow 80% conservative and 20% liberal accounts, and vice versa for liberal agents.

Table 2: Treatment Groups

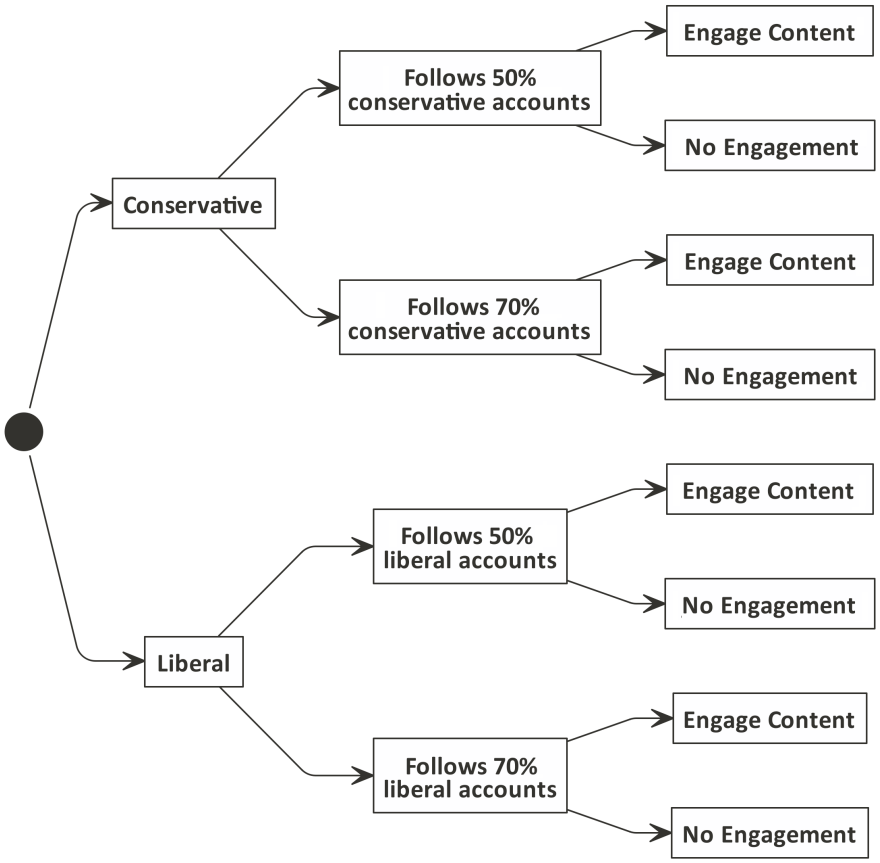
condition	ideology	accounts	sessions	tweets
Control	Liberal	19	871	92990
Control	Conservative	22	906	96700
Only Explicit	Liberal	19	1065	113244
Only Explicit	Conservative	25	1011	107854
Only Implicit	Liberal	18	800	85668
Only Implicit	Conservative	20	707	75458
Explicit and Implicit	Liberal	18	857	91364
Explicit and Implicit	Conservative	29	1161	124064

For the experiment, we gathered two pools containing 600 liberal and 600 conservative twitter accounts. The pools consisted of members of congress, political commentators and politically-sided media outlets. For each agent, we used this pool to draw 200 accounts according to the explicit treatment assignment. Agents, who were not assigned to the explicit treatment were supposed not to provide any explicit

information on their ideology. Now, because Twitter requires you to follow accounts to have anything appear on your timeline these accounts needed to follow some accounts to get started. Consequently, since an agent that follows no accounts cannot receive content, the control group for the explicit treatment consists of neutral explicit cues (50% liberals and 50% conservative).

Implicit cues consist of behaviors informing about a users' preferences, without the explicit objective of selecting content. On Twitter, this can include scrolling rhythm (pausing longer on congruent content or scrolling faster on counter-attitudinal content), clicking on links, like, retweeting, etc. For the sake of technical feasibility, we chose to focus on two implicit behaviors: liking and retweeting. These two behaviors were not only easy to implement, but are also essential to the Twitter community. Consequently, agents assigned to the implicit treatment had a 30% probability to retweet or like congruent content. Agents who were not assigned to the implicit treatment simply scrolled down without interacting at all with content, thus providing no implicit cues.

Figure 11: Experimental Design



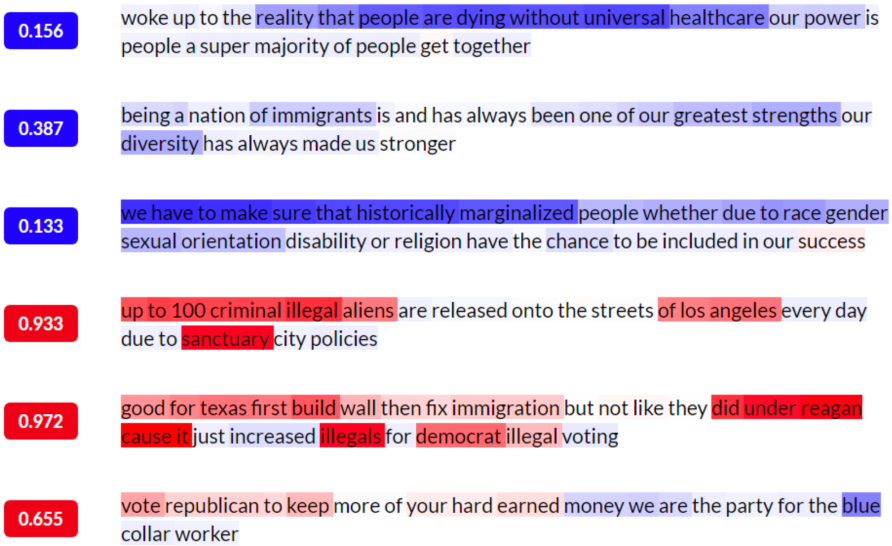
In summary, we adopt a factorial experimental design with three branches, as presented in Fig. 11. The first treatment is the ideological position of the agent (Liberal vs. Conservative) ; the second treatment is in regards to the administration

of explicit cues (following 80% congruent accounts vs. following 50% congruent accounts) and the last treatment regards implicit cues (liking and retweeting congruent content vs. no interaction at all).

3.3.2 Measuring Ideology

Both the implementation of this design, as well as the analysis of the results require an automated and robust strategy to measure content ideology. Because of the implicit treatment, we need to measure the ideology of each individual tweet as the agent browses Twitter. This is even more challenging and excludes hand-coding account-level or other estimation procedures at the account-level as proposed by Barberá 2015. Following Gottlieb 2018, we used supervised machine learning and train a classifier to predict the ideology of a tweet given its text. The classifier was trained on all available tweets produced by acting members of congress (300 000 tweets). Each tweet was labeled as 'Liberal' or 'Conservative' depending on the party affiliation of its author. The neural network combines convolutional layers with a LSTM layer, which has proved to deliver excellent predicting performance in the context of text-classification (Goldberg 2016). The output layer relies on a sigmoid activation function that fits the binary structure of the target.

Figure 12: Qualitatively sampled Tweets with predicted Ideology Score



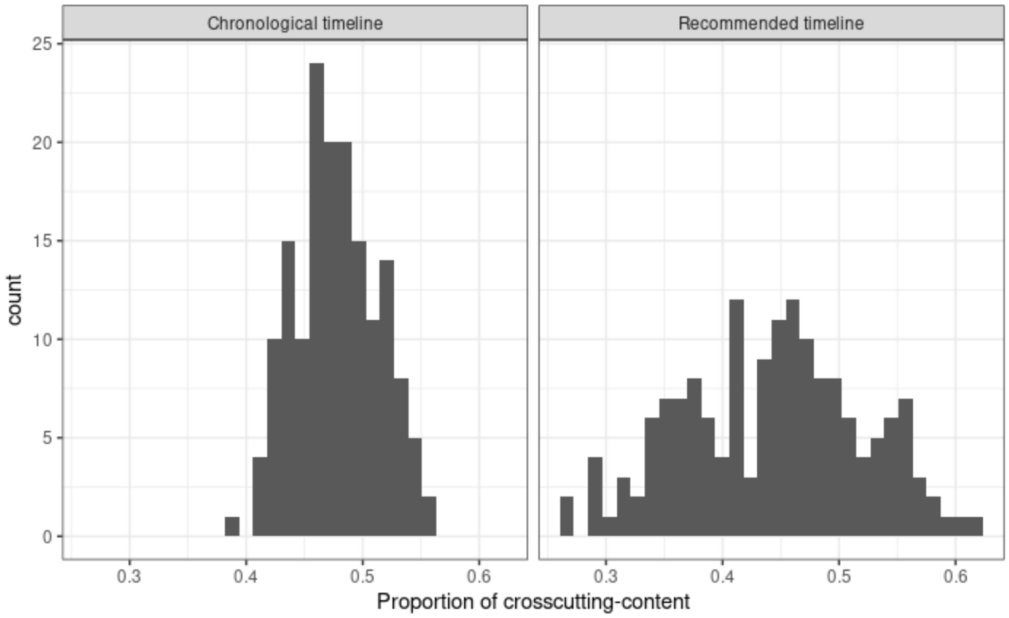
To evaluate the model, we initially kept the tweets of 20% (60 000 tweets) of the congress members aside as a hold-out dataset. We assess the performance of the model with these tweets, which are unknown to the model (as can be seen in Fig. 18). After tuning a basic set of hyperparameters, we reached an out-of-sample accuracy of 86% on content level. To ensure the calibration of the model on the tweets seen by the agents during the experiment, we drew a sample of 1000 random

tweets and ad hoc coded them. Our hand-coding agreed 84.2% of the time with the classifier.

3.3.3 Recommendation Bias

We measure recommendation bias at the level of timeline. Each timeline consists in a sequence of 50 tweets. Recommendation bias is defined as the ideological difference between the recommended and the chronological timeline. Comparing the algorithmic timeline with the chronological ones allows control for any self-selective exposure and volume effect. If we observe that the timeline of a liberal user is overwhelmingly liberal, it is tempting to conclude that the algorithm prioritizes liberal content according to these users' preferences. But this result could be independent from the algorithmic recommendation if the user follows more liberal accounts or simply if liberal accounts produce more content. When comparing the content as recommended with the content as it would appear chronologically, we can control for both self-selective exposure and volume effect, hence isolating the mere influence of the recommendation on the timeline. We assume here that the chronological timeline best represents the unbiased timeline, as it consists of a time free from any algorithmic influence. In our approach, bias does not relate to the preference of a user but to the raw unstructured content, as pre-selected by the user when following accounts.

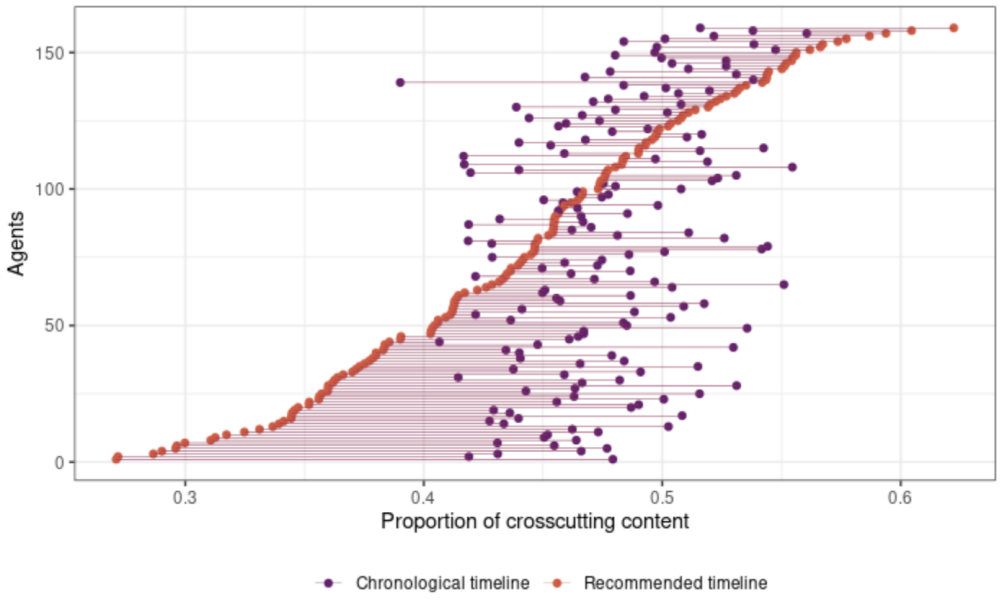
Figure 13: Measurement Distribution: Proportion of crosscutting content in chronological and recommended timelines



Capturing recommendation bias consequently requires observing both the algorithmic timeline and the chronological timeline. An algorithmic timeline consists

of the first 50 tweets appearing on an agents' timeline after log-in. Chronological timelines were artificially reconstructed. During the experiment, we gather all the tweets produced and shared by the accounts, who were followed by the agents. Using the information on (1) which account was followed by an agent and (2) when the agent logged-in to gather the algorithmic timeline, we reconstructed what this agent would have seen, if its timeline was structured chronologically. To make it comparable to the algorithmic timelines, we restricted the chronological timelines to 50 tweets. Using this technique, we were able to match each algorithmic timeline with its chronological counterpart.

Figure 14: Measurement of Recommender Bias: Each vertical line connecting two points represents an automated social media bot. The difference between the recommended (red) and chronological (purple) dot is the observed Recommender Bias in a bot's timeline



Once we have matched each collected algorithmic timeline with its counterpart, we compare the proportion of cross-cutting content entailed in the two timelines. Using the above-mentioned classifier, we label each observed tweet as liberal or conservative. For each agent, we can, hence, compute the proportion of counter-attitudinal (cross-cutting) tweets observed in the algorithmic and chronological timelines. Recommending bias is consequently estimated through the difference between these two proportions. For instance, if the chronological timeline of a liberal agent entailed 40% cross-cutting - conservative in this case - content and its algorithmic timeline contained 30% cross-cutting tweets, we estimate a recommending bias of 10%. This means that for this agent, algorithmic timelines entailed on average 10% less cross-cutting tweets than in its chronological timeline.

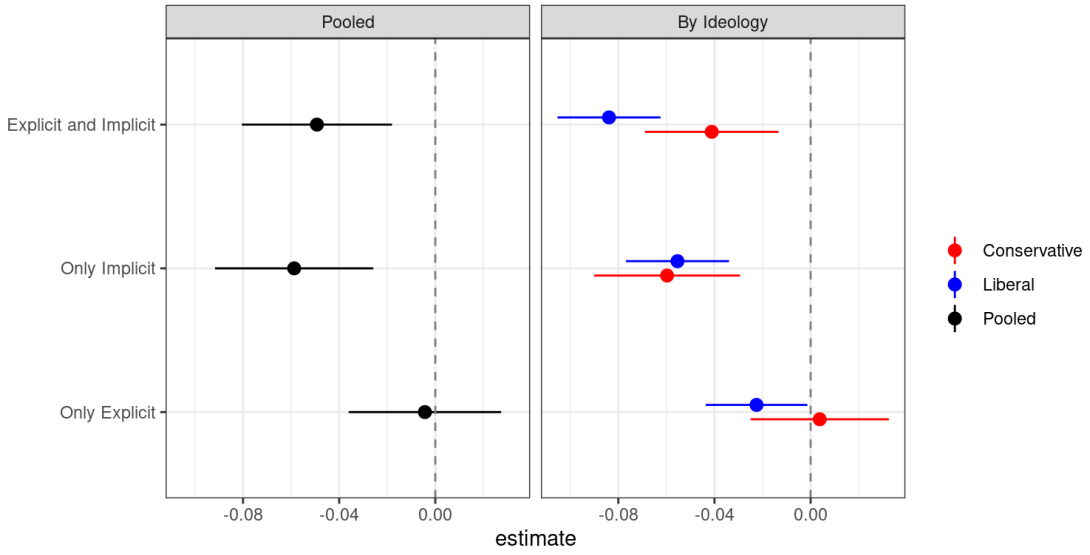
Following our factorial design with three levels, we measure the causal effect of

implicit and explicit cues on the recommending bias, captured as the average difference in proportion of crosscutting content in the chronological and the algorithmic timelines. Treatment assignment was fully randomized, so that we report simple average treatment effects (ATEs). Each treatment branch is compared to the control group within a linear regression (OLS).

3.4 Results

We are mainly interested in the average treatment effect, respectively the average recommendation bias per condition. First we averaged the predictions per timeline, session and bot, as the latter is the unit of analysis. For the sake of clarity Figure 15 presents the main coefficients of interest (Complete regression table can be found in Appendix). We estimated three different models: one with all agents, one with only the liberal agents and one with only the conservative agents.

Figure 15: ATE for Crosscutting Content by Condition



The left hand side shows the pooled estimates over conservatives and liberals for our treatment conditions, while the right hand side shows the conditional treatment effects (CATE) for Liberals and Conservatives. Compared to the chronological timeline and the control group we do not observe any significant effect regarding the explicit cues only branch. When users only follow congruent accounts without interacting with the content, the proportion of crosscutting content is similar in the recommended timeline as in the chronological one.

In contrast, we can observe a strong negative effect for implicit cues. Liking or retweeting the algorithm reduces on average the proportion of cross-cutting content by 6%, even if a user follows as many conservative as liberal accounts. When mixing explicit and implicit cues, we observe a different pattern for liberal and conservative

agents. As explicit and implicit cues interact positively for liberal agents, we do not observe such interaction for conservative agents. This suggests that following more liberal accounts is interpreted as a cue by the algorithm, while following more conservative accounts does not make any difference. This differentiated effect should be interpreted carefully as the confidence interval between the only implicit and explicit X implicit treatments overlap.

3.5 Conclusion

In this study, we conducted an experiment to estimate the ideological bias of Twitter's recommender system. We use automated agents to treat Twitter's algorithm with information cues regarding the ideology of agents. Agents were randomly assigned to an ideology and to a set of informative cues. They revealed their ideology by following a majority of congruent content producers (explicit cues about content preference) and/or by interacting with congruent content (implicit cues about content preference). The experiment took place during the 2020 presidential election.

First, we find evidence for the filter bubble hypothesis. After three weeks of treating the algorithm with ideological cues, we observe up to 8% less cross-cutting content compared to a chronological timeline. This estimate is likely to represent a lower bound, as our experiment only lasted three weeks. In real life, users share ideological cues for a much longer period, so that much bigger effects can be expected.

We further find that the Twitter recommender system responds differently to different types. More precisely, the algorithm responded much heavier to agents directly interacting with the content compared to agents simply following user accounts. This suggests that user interactions play a major role for the Twitter algorithm when it comes to learning preferences of their users. This has important implications because users might not even be aware of this strong personalization effect when they like or share content. On the other hand, the act of following congruent accounts did not result in any significant effect on the ideological slant of the timeline. Again, the effect might be underestimated, as in real life the proportion of congruent accounts followed by a politicized person is likely to be much higher than 80%. The results of this study are relevant for our understanding of online polarization. We have demonstrated that it is possible for a user to receive an ideologically slanted timeline, even if she explicitly follows ideologically diverse accounts to receive content from both sides of the political spectrum. This also suggests that looking at who is followed is not enough to understand online ideological diets. Because of personalization, the content that appears in a person's timeline is not necessarily representative of the overall content produced by the following account.

Unknowingly consuming personalized content can also bias citizens' perception

of their environment. This can create a dissonance between the reality as perceived online and as described in the media, which could possibly lead to institutional mistrust. More research is needed on the exact role played by the awareness of personalization. As other social media also rely on personalization, this issue will remain relevant and important for the foreseeable future.

Finally, we showed in this study that recommender systems can heavily impact the content consumed by users. In this context, social media platforms carry the responsibility of improving the transparency around their recommender system. This includes both direct communication with users as well as the creation of research opportunities enabling independent research to properly scrutinize recommender systems. Setting up this study was a herculean task and expanding it to other types of cues for a longer period seems impossible and will not be possible without the active help of social media companies.

4 Does Communication Signal Voter Preferences and Participation?

Sascha Göbel and Simon Roth

Abstract

Does communication carry information about voter preferences and participation? This question is central to explanations of communication effects on vote choice and turnout, language use in politics, and communication-based measures of electoral characteristics. Yet empirical evidence in this regard is scarce. Existing work typically assumes communicated information from discussant characteristics, focuses on linguistic differences in political orientation, and investigates elite or political talk. In this paper, we examine whether the full content of exclusively citizen communication signals their partisan preferences and electoral engagement. We draw on a unique dataset that combines US voter registration records with communication on Twitter. To explore communicative signals, we rely on a neural network approach to natural language processing. We find that US citizens' talk conveys substantive information about their party preference and is instructive about their participation at low-stimulus elections and thus high voter engagement. Predictive signals suggest that political identities fundamentally structure communication.

4.1 Introduction

Does communication signal voter preferences and participation? Extant literature demonstrates significant effects of interpersonal communication on vote choice and turnout (Beck et al. 2002; Eveland & Hively 2009; Huckfeldt & Sprague 1991; McClurg 2006). The notion that communication content transmits individual voter preferences and participation to others provides a central explanation for these effects. Information about one's party affiliation, sense of civic duty, or political efficacy can serve others as voting cues or incentives to participate. Yet empirical evidence on whether interpersonal communication actually signals such information is lacking. Existing, largely survey-based studies are not equipped with respondents' specific communication records (Schmitt-Beck & Partheymüller 2016). Communicated information is hence typically assumed from discussant characteristics rather than inferred from actual messages.

Evidence on whether communication reflects voter preferences and participation is not only relevant to the assumptions underlying the relationship between interpersonal communication and electoral characteristics, though. It also informs language use in politics. The pertinent literature focuses on specific linguistic differences along political orientations (Okdie & Rempala 2019; Sterling et al. 2020). Considering variation in communication at large offers a more comprehensive perspective with insights on the substance of political traits. To the extent that communication carries information about voter preferences and participation it may also prove relevant for measurement purposes. Interpersonal communication is available in plenty on social networking platforms. Computer scientists have explored this source extensively in this regard. However, this rather technical work focuses on self-reported political orientation, is usually based on political rather than ordinary talk, confounds elite and citizen communication, and concentrates on specific vocabularies or linguistic features (Hinds & Joinson 2018).

This paper adds to the existing literature by investigating whether the full content of exclusively citizen communication signals their partisan preferences and electoral engagement. To motivate and contextualize our empirical analysis, we advance a theory about why communication should mirror voter preferences and participation in the first place. Starting from social identity and self-categorization theory (Tajfel 1981; Turner et al. 1987), we posit that voting and partisan attachment provide foundations for social identification (Huddy 2001; Kalin & Sambanis 2018). We argue that group norms that constitute the resulting political identities influence communicative behavior (Hogg & Reid 2006). Consequently, individuals' perceived identities should crystalize in communication and reflect their partisanship and electoral engagement. We discuss research on affective polarization and social norms of

voting to clarify communicative behaviors, which we expect to inform voter preferences and participation.

To circumvent data limitations in prior studies we draw on a unique dataset that combines individual-level administrative information with digitally recorded communication for over 36,000 voters. Specifically, we use data by Göbel 2020 that merges public voter registration records from the state of Florida with corresponding public Twitter accounts and that we expand by data from five additional US states. These data allow us to operationalize both our outcomes of interest, voter preferences and participation, based on voters' official party registration and validated turnout at the 2018 general and primary elections. Simultaneously, more than 11 million text messages recorded on the Twitter accounts of these citizens and ranging between 2006 and 2020 grant us direct access to communication content at the level of the message rather than its source. Importantly, we deem the US a "likely case": Political identities are highly developed and salient in the US. Considering our theoretical background, communicative signals of voter preferences and participation should be particularly prevalent here.

To explore unknown patterns in communication that speak to voters' partisan preference and electoral engagement we rely on predictive modeling (Cranmer & Desmarais 2017). We develop an artificial neural network that is theoretically informed by an application to natural language and text data (Chollet & Allaire 2013). This approach allows us to consider the full spectrum of communication, including a vocabulary of over 600,000 features as well as the semantic relationship, context, position, and order of words. We train and evaluate our model using repeated K-fold cross-validation. This way we are able to assess the degree and variability to which communication generally signals our outcomes of interest. Additionally, we employ Local Interpretable Model-agnostic Explanations (Ribeiro et al. 2016b) to offer substantive insights on what citizens are saying that resonates with their partisan preference and electoral engagement.

We find that communication signals voter preferences and participation, although to varying degrees. Individuals' communication is highly predictive of their party preferences, offering a rich source of information in this regard. Communication is also moderately instructive about voter participation, though only as regards low-stimulus elections. This suggests that communication carries information about the voting habits of likely voters but not the electorate in general. Communicative signals that are predictive of voter preferences and participation reflect group norms that constitute political identities and show a strong connection between partisan identity and voter engagement.

In the next subsection, we situate our study among related work on interpersonal communication and electoral characteristics, the use of language in politics, and the

measurement of political traits and carve out its contribution. We then discuss our data, case selection, and methodological approach, before turning to our empirical findings. We conclude with a discussion of our study's limitations and consider its implications with reference to our theoretical background and related literature.

4.2 Related Work

Interpersonal communication occupies a central role in the study of voter preferences and participation. The question whether communication influences these individual electoral characteristics has received particular attention. A large strand of literature has offered evidence that interpersonal communication significantly influences individual party preferences and vote choice (Beck et al. 2002; Huckfeldt & Sprague 1991; Kenny 1998; Schmitt-Beck & Partheymüller 2016). Similarly, several studies highlight the positive effects of interpersonal communication on voter participation (Carlson et al. 2020; Eveland & Hively 2009; McClurg 2006).

A common feature among these survey-based studies is that they measure communication content through discussant characteristics rather than via the messages as such. Effects of interpersonal communication are assessed by way of discussant network composition and discussion frequency. In consequence, actual communication and its content has only played a subordinate role in the existing literature (Schmitt-Beck & Partheymüller 2016). At the same time, the notion that communication content transmits discussant characteristics to others is at the heart of this research, as it provides an important explanation for the effects of interpersonal communication. Note that these effects by themselves, however, do not prove that communication content carries discussant characteristics. Unobserved confounders that influence both discussant characteristics and the interpersonal network may just as well explain such findings. That communication reveals, for example, individuals' party preferences, sense of civic duty, social norm of voting, or political efficacy, which can act as cues or incentives to participate, is generally assumed. By focusing on whether communication signals voter preferences and participation, we shed some light on this thus far unstudied assumption.

A strand of literature that considers actual communication with reference to political characteristics is the study of political language. This literature is largely involved with identifying linguistic differences along political orientations. Most of the studies focus on political elites. Cichocka et al. 2016 and Diermeier et al. 2011, for instance, investigate differences in the use of grammar between liberal and conservative politicians and Slapin & Kirkland 2020 study pronoun usage, sentiment, and complexity of speeches of politicians dissenting from the party line. Only recently have scholars conducted similar analyses with reference to citizens. For

instance, Okdie & Rempala 2019 and Sterling et al. 2020 study language differences between liberal and conservative undergraduate psychology students and Twitter users across several linguistic categories.

These studies, although having made invaluable contributions to our understanding of language use in politics, are limited by an exclusive focus on political orientation and a narrow perspective at communication that often focuses on psychological aspects. Analyses are usually restricted to very specific features of language, frequently identified via the Linguistic Inquiry and Word Count dictionary.[5] This is paired with a commonly used bag of words approach that considers only the frequency of words in documents, ignoring other informative properties such as semantic relationships, context, and structure. This disregards whether communication in its entirety reflects political characteristics, such as voter preferences and participation, likely accounting for the typically small effects in this literature (Sterling et al. 2020). We overcome this limitation by using a neural network approach to natural language processing. In doing so, we are, to the best of our knowledge, the first to study the full spectrum of communication as it refers to voter preferences and participation.

Another burgeoning literature treats voter preferences and participation as traits that can be predicted from behavioral trace data. Barberá 2015 uses Twitter follower networks and Kosinski et al. 2013 use Facebook likes to measure social media users' party affiliation. Bach et al. 2020 rely on survey respondents' web browsing records to predict their political views and electoral participation. Beauchamp 2017 uses text messages on Twitter to measure vote intentions but only aggregated at the level of states. At the individual level, actual communication on social media has received vast attention in computer science. However, in this literature, the focus is less on the message and information transmitted by communication. It rather emphasizes technical aspects in search of the best performing textual classifier of political orientation. As a result, this research has put a focus on political instead of general communication, has mixed communication by political elites and citizen,[6] and has employed limited vocabularies and specific linguistic features (Hinds & Joinson 2018). Instead, our data and methodological approach offer insights on whether (exclusively) citizens complete and everyday communication on Twitter yields a viable alternative for measuring individual political preferences and participation – information that is generally unknown for the social media population. Therefore, our study also adds to this literature.

In sum, we add to the existing literature in three ways. First, we inform a com-

[5]Diermeier et al. 2011 present an exception, since they also focus on how the content of political speech reflects ideology, though only with view to political elites and within the limits of the bag of words approach.

[6]This is one important reason why we excluded so called retweets in our empirical analyses.

mon but unstudied assumption in the literature on communication effects on vote choice and turnout. Second, rather than focusing on highly simplified representations of language and specific linguistic categories, we consider whether communication in its entirety reflects differences in political views and electoral engagement. Third, we sound out the use of citizens' social media-based communication for measuring political traits.

4.3 Theory

We start our empirical discussion with a theory about why communication should mirror voter preferences and participation in the first place. Considering that extant literature commonly treats this as an automatic side effect of political talk, we deem theorizing in this regard particularly important. Moreover, the theoretical scaffold aids in understanding and contextualizing our empirical findings.

Communication theory suggests that communicative behavior is structured by groups (Hogg & Reid 2006). Common thoughts, feelings, and behaviors within groups, aka group norms, have a bearing on what people say. Group norms become directly or indirectly the subject of talk. For instance, to gain acceptance while socializing with others, to justify actions in group decision-making, while discussing opinions on media reports, parenting, or the behavior of others, people express what is and what is not normative of their group, thus delineating group boundaries. In consequence, communication carries information about individuals' perceived group membership.

This line of reasoning is rooted in social identity and categorization theory, which explains behavior on the basis of identification with groups (Tajfel 1981; Turner et al. 1987). This school of thought maintains that individuals categorize themselves and others in social groupings. Context and socialization determine the social category that persons sort themselves into based on common attributes and lay out norms and appropriate behaviors. Over time, people develop a sense of belonging to the social grouping – a perceived, subjective group membership[7] – and make their behavior conform to respective group norms. Throughout this process of identification, the perceived group membership is internalized, becoming part of one's self-concept. Such social identities, once developed, have behavioral implications. To preserve consistency with their self-concept and avert psychological discomfort, people strive for positive distinctiveness in comparison with others. This motivates their social identity and harvests its benefits as a source of positive self-esteem (Monroe et al. 2000). Accordingly, persons focus on and highlight differences between social groupings, naturally placing their social category favorably over others. From a

[7]This differs from objective group membership in that it is subject to individual discretion and does not require formal membership

communication theoretical perspective, these consequences of social identification are observable in communicative behavior.

Classic social identities typically relate to non-political groupings such as ethnicity, religion, sexuality, or class. However, political science scholars have highlighted that political categories provide a basis for social identification as well (Huddy 2001; Kalin & Sambanis 2018). This is plausible considering the interdependent, group-based nature of political behaviors (Huckfeldt 2005). According to Huddy 2013, when group norms pertain, for example, to specific political parties or courses of political action, they form political identities.

Recent research shows that voter preferences map to coarse categories that function to distinguish parties on the same side from those at the opposing side (Bolstad & Dinas 2017). Voters were shown to consciously sort into and identify with specific partisan groups and act more favorably towards them compared to others. Especially in the US context, scholars have repeatedly shown that partisanship represents a profound political identity (Greene 1999; Theodoridis 2017). Egan 2020 reports evidence that Americans even shift other social identities to align with their party identification. The behavioral consequences of partisan identification are best demonstrated by the literature on affective polarization in the American electorate (Iyengar & Westwood 2015; Mason 2015). Democrats and Republicans face each other with prejudice and animosity. Identification with a party increasingly means intolerance, dislike, and distrust for the other side. Partisan bias colors not only voters' behavior toward each other, but also their judgment of actions and beliefs about facts. We expect these behaviors to mirror communication and reveal voter preferences.

Electoral participation is strongly related to party identification. Voting allows for substantiating one's perceived party affiliation (Dinas 2014) and for promoting positive political outcomes for the own group (Fowler & Kam 2007). Supporting this, Huddy et al. 2015 find that those who identify strongly with a party are also more active at the polls. By implication, the behavioral consequences of partisan identification should also inform voter participation.

However, voter participation may also function as a political identity in its own right. This is most clearly expressed in the debate on effects of priming a social identity as a voter on turnout (Bryan et al. 2019; Gerber et al. 2018). Voting is a group-based activity (Bhatti et al. 2020) and subject to the degree of social approval (Fieldhouse & Cutts 2020) or disapproval (Partheymüller & Schmitt-Beck 2012 in one's social context. The essential group norm is either the shared belief that most people in the group vote, ought to vote, and expect others to vote or the inverse. The degree of political efficacy and whether voting is considered a civic duty within the social context is central to this idea (Blais & Achen 2019). We would expect such

ethical concerns to show in communicative behavior and reflect voter participation in particular.

4.4 Data and Methods

Communication content is distinctly lacking from the largely survey-based research on the relationship between interpersonal communication and individual electoral characteristics. Surveys, while well suited for the collection of individual attributes, are not designed for recording respondents' actual communication, let alone at large scale (Schmitt-Beck & Partheymüller 2016). We believe that these data limitations account for why our research question, despite its obvious relevance for the existing literature, has thus far remained unstudied.

To overcome this deficiency, we use data from Göbel 2020 that combines public voter records of a large sample of registered voters from Florida with their public Twitter accounts. These data uniquely allow for observing persons' official party registration and validated turnout together with their actual communication content. Originally, the two data sources were integrated via a custom algorithm using publicly available email addresses in the voter records. To extend this data both in size and geographically, we collected further voter records for the states of Arkansas, Colorado, Connecticut, Rhode Island, and South Dakota for which we were able to access public contact information. We identified respective Twitter accounts by pairing this information with integration services by the identity resolution company FullContact.

We operationalize our outcomes of interest, voter preferences and participation, using official validated information from the voter records. When registering to vote, voters in the US are given the choice to declare an affiliation with a political party. In closed primary states, this party registration is required for voters to participate in partisan primary elections. That said, recent research shows that primary voters are generally representative of a party's electorate (Sides et al. 2020) and that crossover voters, those who register with the opposing party to vote for an inferior candidate in the primary, are rare (Norrander 2018). Research has also repeatedly demonstrated a strong link between party registration and individual partisan attachment (Burden & Greene 2000; Thornburg 2014). Consequently, we rely on party registration as a binary measure of voters' party preference as either Republican or Democratic. As regards voter participation, we focus on the 2018 general and primary elections. These are the latest elections for which validated individual turnout indicators are available in our data. We use these binary indicators for whether a person voted or abstained in the respective election, conditional on eligibility, to measure individual voter participation.

To operationalize communication, our input of interest, we rely on text messages communicated through the Twitter accounts of registered voters in our sample. Twitter is a social networking service that provides a major venue for people to communicate, interact, and take part in political discourse (Nagler & J. Tucker 2015). We focus exclusively on text messages sent directly by voters in our sample and exclude shared messages, so called retweets. This allows for measuring the content of communication at the level of the message and its source.[8] Public text messages sent by registered voters in our sample were collected using Twitter's REST API. The data by Göbel 2020 includes daily records of text messages since August 2018. For our extension of the data, the daily collection commenced in December 2019. However, since Twitter profiles store up to 3,200 past text messages, our data additionally includes communication that goes as far back as October 2006. Ranging until July 2020, we record almost up to 14 years of social media-based communication content per person. The final number of voters and Twitter messages in our data across sources and outcomes is shown in Table 3. Further details on data collection, integration, and processing are given in subsection C.1 of the Supplementary Information (SI).

While our focus on the US is evidently driven by data requirements, it is also theoretically appealing. Political social identities are highly developed and salient in the US, possibly even more so than in any other country. Seen against our theoretical background, if communication does not signal voter preferences and participation here, it is less likely to do so in countries where political social identities are less important.

Table 3: Voters and Text Messages in the Dataset

	Party Preference			General & Primary Election Turnout		
Voter File	Voters	Messages	Batches	Voters	Messages	Batches
Arkansas (Sep-2019)	932	184,276	9,104	932	184,276	9,104
Colorado (Aug-2019)	1,602	309,743	14,441	—	—	—
Connecticut (Oct-2019)	1,634	345,975	16,622	1,634	345,975	16,622
Florida (Oct-2017)	23,122	7,631,653	367,495	31,964	10,517,440	509,346
Rhode Island (Sep-2019)	1,629	369,551	17,968	1,629	369,551	17,968
South Dakota (Sep-2019)	125	24,298	1,299	159	31,528	1,622
Total	29,044	8,865,496	426,929	36,318	11,448,770	554,662

Note: The columns 'Voters' list registered voters for whom outcomes and text messages are available. The columns 'Messages' list the number of single text messages after text pre-processing. The columns 'Batches' refer to text messages stacked to sequences of a maximum length of 300 words. Dates of origin for the voter records are indicated in brackets. Turnout indicators are not publicly available for registered voters in Colorado.

[8] Another reason for the exclusion of retweets is that they may introduce leakage from the train into test data if persons in both datasets share the same text message.

Recall that our goal is to detect signals in communication that relate to voter preferences and participation. This search for unknown patterns is an exploratory task, for which predictive modeling is ideally suited (Cranmer & Desmarais 2017). At the heart of predictive modeling is so called training data, input data for which associated outcomes of interest are already coded. Based on these data, a model is employed to distill the input features that are most predictive of distinct outcomes. To assess whether the identified features are generally predictive of and thus signal the outcome, the trained model is applied to new, previously unseen data.

Dealing with written communication and hence text data, we require a predictive model that can interpret and process natural language. As pointed out above, existing studies mostly rely on approaches that represent text as a (typically sparse) matrix of word counts in documents, aka a bag of words. This has several limitations. In our context, facing millions of texts with a massive vocabulary, computational constraints present the biggest obstacle. Scholars routinely cut their data down to accommodate the computational costs of representing text as a matrix. Furthermore, this text interpretation does not incorporate the order, the context, and the semantic relationship of words.

Rather than making compromises on all these dimensions, we rely on a neural network approach to natural language processing (Chollet & Allaire 2013). In neural network models, words are mapped to unique integers and text inputs are represented as sequences of these integers. In addition to substantially reducing the dimensionality compared to matrix representations, this vector format also retains more information, such as the position and context of words in a sequence. This provides a more complete interpretation of communication and allows us to make full use of our data. Instead of pre-selecting a limited range of features, we let the data speak about whether and which features are predictive of our outcomes. Before describing our neural network model together with its training and evaluation, we turn to the text pre-processing that we applied in accordance with the requirements of neural network models.

4.4.1 Text Pre-Processing

Prior to transforming communication into machine readable format, we performed limited but necessary processing operations on the raw text. In line with our interest in written natural language, we first removed emojis, emoticons, and URLs. To assist the neural network model in picking up relevant patterns, we further removed very infrequent features that occurred in less than three messages. Their scarcity makes them 'noise' that provides room for the model to pick up message-specific peculiarities rather than general differences between groups.

Similarly, we applied basic text normalization to align words into common forms.

We replaced contractions with their long form and elongated words with their standard form, removed kerning and punctuation, and finally lowercase all words. However, to retain semantic differences between words, we did not apply stemming. We also did not remove stop words, which can provide valuable context information (Chang & Masterson 2020). To distinguish between the actual message, its topical emphasis, and references to other Twitter accounts, we kept hashtags and mentions.

Twitter is a social networking platform that operates in the microblogging domain. Accordingly, messages are brief and can range anywhere from just 1 up to 280 (formerly 140) characters. Single messages in our data are on average just 16 words long. This paucity of information in a text sequence associated with a specific outcome can make it difficult for the model to detect relevant patterns in communication. We also wouldn't expect every single message to contain information that speaks to voter preferences and participation. For this reason, we stack all messages of a voter in order of their creation and split the resulting text into fixed sequences of maximum 300 words. This provides more information to the model and a somewhat more representative summary of persons' communication than one single message. Indeed, Figure 25 in the appendix shows that stacking texts to longer sequences helps the model substantially in detecting relevant signals in communication.

Finally, we bring the text sequences into a format that can be processed by the neural network. This involves arbitrarily assigning a unique integer identifier to every feature in the vocabulary, a process called tokenization. The tokenized vocabulary is then used to transform the texts into integer-encoded sequences. As the model expects sequences to have the same length, shorter sequences are padded with 0s. For example, "undocumented immigrants should be given health insurance" turns into [5, 1, 3, 9, 7, 6, 2] and "illegals should be deported" into [0, 0, 0, 8, 3, 9, 4]. Depending on the outcome, these processing steps leave us with an immense vocabulary of 615,869 features pertaining to 490,796 text sequences on average (see the columns 'Batches' in Table 3 for a breakdown across outcomes and states).

4.4.2 The Model

Neural networks take text sequences as input and operate on their elements through multiple processing layers. Considering just one text sequence for simplicity, a general neural network is represented by the following forward recursive formula (Calin 2020):

$$x_j^{(l)} = \phi\left(\sum_{i=1}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)} - b_j^l\right) \qquad \text{with } \; 1 \leq l \leq L. \tag{1}$$

A neuron, the building block of a neural network, is represented by $x_j^{(l)}$, whereby j denotes a neuron in the current and i a neuron in the prior layer. Layers of the

network are indexed by l, $l = 0$ describes the input layer, $l = 1$ the first hidden layer, and L the final output layer. The number of neurons in a layer is given by $d^{(l)}$. At the input layer, $l = 0$, neurons in the vector $x^{(0)}$ are equivalent to integer-encoded words in the input sequence, so $d^{(0)} = 300$. The neurons are assigned random weights $w^{(0)}$, which are transformed by an activation function ϕ to range between 0 and 1. The resulting vector enters as $x^{(l-1)}$ into the computation of the neurons at the first hidden layer, $l = 1$. The weights at this layer are also randomly initialized. The weight of the j^{th} neuron w_{ij} connects to the i^{th} neuron in the prior input layer and is hence multiplied by the respective activation $x_i^{(l-1)}$. This is repeated over all neurons in the previous layer with reference to the j^{th} neuron in the current layer, hence a network. The resulting vector is averaged, shifted by a bias b_j^l (initially 0), and transformed by the activation function to return $x_j^{(l)}$. Repeating this process for all remaining neurons in the current layer yields a vector of activations that again serves as input for the next hidden layer, and so on. Ultimately, we aim for the neuron x^L, the output of the final layer. In our case, this is a single number that ranges between 0 and 1 and that describes the probability that the text sequence signals the class of an outcome of interest, i.e., a Republican (0) or Democratic (1) party preference.

After one such forward pass through the network, a cost function is evaluated that measures the distance between the output of the final layer and the actual outcome class associated with the input sequence. Using this feedback signal, a gradient-based optimization algorithm is employed to adjust all the weights and biases in a backward pass through the layers of the network in a way that minimizes the cost function. By propagating nudges to the weights and biases back to the input layer, the model is trained to identify patterns in the input sequence that matter most for predicting the outcome class.

Of course, the actual training procedure involves not just one but many input sequences and associated outcomes. Similarly, the specific architecture of the layers affects the number and dimension of neurons at each stage of the network. This condensed representation serves merely to illustrate the basic logic through which a neural network model identifies signals in communication that speak to voter preferences and participation.

As regards the layer architecture, we designed our model based on recommendations in the literature and theoretical considerations with an application to text data. We start with an embedding layer (Rheault & Cochrane 2020). This layer transforms integer-encoded words in the input sequence into numeric vectors, so called word embeddings. The purpose of this is to capture the semantic relationship between words, whereby the proximity of word vectors informs similar meaning. As with any other layer, the values of the word embeddings are initialized randomly

and adjusted through the process of backpropagation described earlier. Although now incorporating information about word relationships, the layer still treats the words in the input sequence separately.

To capture word context, we employ a one-dimensional convolution layer (Chollet & Allaire 2013). Instead of focusing on each word embedding in isolation, this layer considers local combinations of word embeddings. This allows for detecting feature patterns that are predictive of the outcome. A pooling layer summarizes this process by selecting the most distinctive patches (Calin 2020).

So far, the model is still oblivious to the exact position and order of word embedding patterns in the input sequence. To incorporate the structure of the input sequence and time dependence between words into the model, we introduce a long short-term memory layer (Chang & Masterson 2020). This layer maintains a memory of the values of previous patterns in the input sequence, which is accessed when determining values of the current pattern.

Such a setup is suited to detect highly complex relationships between the text inputs and associated outcomes. Indeed, it manages this so well that the model is likely to even take up peculiarities of individual texts seen during weight adjustment. As a result, rather than identifying general signals of voter preferences and participation, the model overfits and detects the 'noise' that perfectly predicts the outcomes in the training data. We rely on two regularization techniques to reduce such overfitting (Garbin et al. 2020). For one, we employ intermittent dropout layers that randomly reset weights. Second, we include a batch normalization layer that normalizes layer inputs, which makes the network less dependent on initial weights and has regularization effects.

Given the vast amount of data we have at our disposal and the huge associated set of features, we opted for hyperparameters that conform to a narrow model (fewer neurons in hidden layers). This serves to keep the capacity of the network in check and safeguard against overfitting as well as excessive runtime. While we have not performed any model tuning with the data presented here, we offer comparisons of several layer architecture and hyperparameter choices in the appendix. These comparisons show that changes to the model do not considerably alter our results.

4.4.3 Training and Evaluation

Neural network models are subject to two sources of variability. One is sampling variability due to specifics of the data used for training and testing the model. The other is model-based and stems from convergence to different local optima during training. To quantify this variability and evaluate our model reliably we make use of repeated K-fold cross-validation.

This entails randomly splitting the available data into K folds of equal size. The

model is then trained on all except the k^{th} fold, which is held out to evaluate the trained model. This is repeated for all $k = 1, ..., K$ folds. To balance bias and variance, which decrease diametrically with the amount of training and test data, we set K to 10 (Kuhn & K. Johnson 2013. To approximate a sampling distribution of model evaluations, we repeat this process 10 times, yielding 100 distinctly trained and evaluated models for each outcome.[9] In order to ensure that the held out test folds are truly out-of-sample, we split the data along voters instead of texts. This way neither text sequences nor voter-specific speech patterns will appear in both the train and the test data at the same time.

We want our model to grasp signals in communication with reference to both classes in our binary outcomes without prioritizing one over the other. For this reason, we balance classes in the training data by randomly downsampling text sequences and associated outcomes of the more common class. To provide neutral ground for evaluating our models and ease interpretation, we also balance classes in the test folds using the same procedure.[10] On average, this leaves us with $348,474$ [95% CI: $231,171$; $465,777$] text sequences and associated outcomes in the training, and $38,720$ [$23,450$; $53,988$] in the test data across outcomes and model runs.

4.5 Results

Our model's out-of-sample classification performance is directly informative about the degree to which communication signals voter preferences and participation. We consider several evaluation metrics to gauge model performance.

Classification accuracy, probably the most common metric, depicts the proportion of outcomes associated with text sequences that are correctly predicted by the model. Precision and Recall offer a more nuanced view. Precision shows the proportion of predictions of the relevant class that actually belong to this class. Recall, in turn, displays the proportion of correct predictions of the relevant class among all relevant classes in the data. The term "relevant" only marks a point of reference. Following Kuhn & K. Johnson 2013, we always take the first level as the relevant class.[11] The F1 score considers both metrics and represents their harmonic mean.

[9] We capped the number of models at 100 due to an average runtime across outcomes of 33 minutes per model.

[10] Neunhoeffer & Sternberg 2019 note that this does not yield a measure of the true predictive performance of a model in the wild, where classes are always imbalanced. We agree. However, our primary goal is not to deploy a classifier for the measurement of voter preferences and participation in the wild, but to assess whether communication in principle signals these characteristics. In addition, evaluating our model on imbalanced test data does not yield a measure of true predictive performance either. This is because, due to the way our data was collected, the distribution of voter preferences and participation does not match their true distribution on Twitter, which is unknown to us. In any case, Figure 29 in the appendix also reports results for evaluating our models on imbalanced data.

[11] For party preference, the first level and relevant class is "Republican" and the second level

A limitation of these metrics is that they are derived from a contingency table of true and false predictions at a single decision threshold. The predicted outcome class of a text sequence is decided by whether the probability returned by the model crosses the conventional 50 percent threshold, yielding the second-level, or not, yielding the first-level class. The receiver operating characteristic curve instead pits the rate of correct relevant-class predictions against the rate of wrong relevant-class predictions at all possible classification thresholds. The area under this curve (ROC-AUC) hence provides a more general, aggregate measure of performance. Similarly, the area under the precision-recall curve (PR-AUC) summarizes how recall and precision relate to each other under varying classification thresholds. The higher these two values, the more communication signals the respective outcome.

Figure 16 presents model evaluation metrics for our three outcomes: party preference, general election turnout, and primary election turnout. The dashed horizontal line at 0.5 represents the benchmark that would be reached by randomly assigning outcomes to text sequences. Metrics falling below this line indicate that communication does not systematically convey information about the respective outcome.

With reference to individuals' party preference, our communication-trained model classifies out-of-sample text sequences with an average accuracy of 72%. The imbalance between precision and recall shows that on average the model wrongly predicts more Republican text sequences as coming from Democrats than vice versa. This suggests that communication picked up by the model as pertaining to Democrats occurs notably among Republican voters. A possible explanation for this is that the extent to which Republican voters are torn over the future of the party and its presidential candidate reflects in communication.

Independent of specific classification thresholds the model records an average performance of 79% ROC-AUC and 80% PR-AUC. The high PR-AUC value shows that the model maintains high recall and precision when increasing the classification threshold, at some point also yielding a balanced prediction of party preferences.

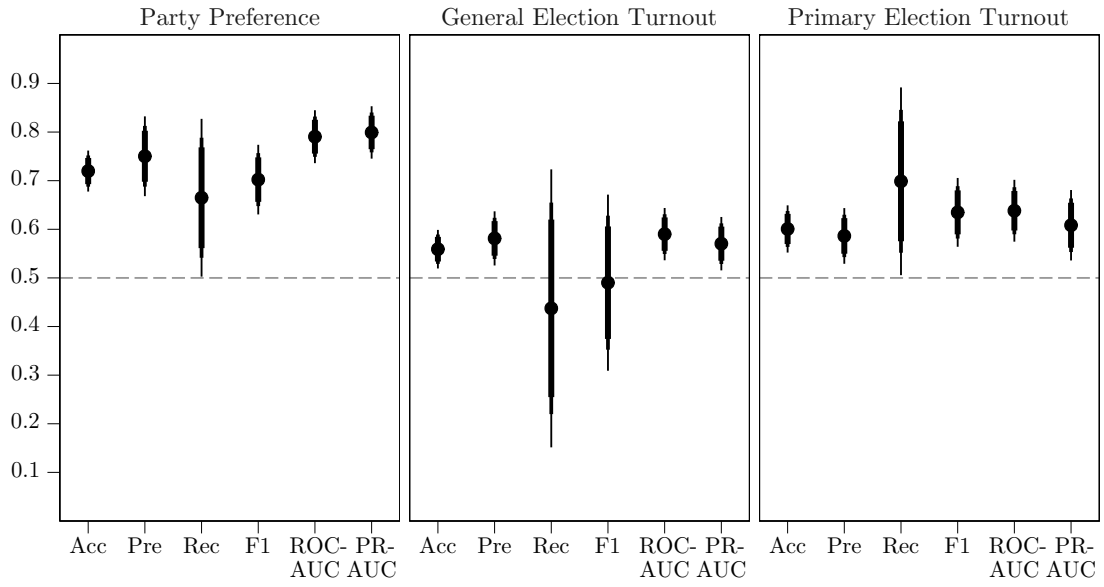
To put these findings into perspective, replicating existing approaches for predicting citizens' political orientation from Twitter communication, Cohen & Ruths 2013 report accuracy values around 65%. The training data that was used in the respective studies, however, focused on communication by Twitter users who publicly state their party affiliation and included communication by political elites as well as retweets. Because of that, prior performance measures are considered overly optimistic as regards signals from exclusively citizen communication (Hinds & Joinson

is "Democrat". For turnout, the first level and relevant class is "abstention/not voted" and the second level is "voted". For example, precision reveals how many of the text sequences predicted as signaling abstention in the primary election are actually associated with abstention and recall shows how many of the text sequences associated with abstention the model managed to capture.

2018).

On that account, our neural network model shows superior classification performance. In addition, the findings emphasize interpersonal communication as a reliable and rich source of information about voter preferences. The communicative signals that our neural network model picks up are highly predictive of Democratic and Republican partisan attachment and none of the metrics' uncertainty intervals cross the 0.5 mark.

Figure 16: Out-of-sample classification performance for predicting party preference and turnout.



Note: Acc = Accuracy, Pre = Precision, Rec = Recall, ROC-AUC = receiver operator characteristic–area under the curve, PR-AUC = precision-recall-area under the curve. Mean and uncertainty intervals (90%, 95%, and 99%) are based on $10{\times}10 - fold cross - validation = 100 models for each outcome$.

Regarding voter participation, the model predicts turnout associated with out-of-sample text sequences with an average accuracy of 56% for the general election and 60% for the primary election. In both cases we see a notable imbalance between precision and recall. For the general election, the model classifies many more non-voter related texts as voter related texts than vice versa. With an average F1 score below and uncertainty intervals overlapping the 0.5 mark, communication does not seem to reliably signal individuals' general election turnout at a conventional classification threshold.

Conversely, for primary election turnout the model sorts more text sequences pertaining to actual voters into the non-voting category than the other way around. Also, all metrics are above and do not cross the 0.5 mark, with the F1 score even at 63%. Note that across classification thresholds communication on average seems informative about both general (ROC-AUC = 59%, PR-AUC = 57%) and primary election turnout (ROC-AUC = 64%, PR-AUC = 61%), though still more so for the

latter.

That precision-recall patterns are reversed between elections and that communicative signals are stronger with reference to primary election turnout is substantively interesting and plausible. General elections are highly salient, attracting even low-propensity voters, whereas low-stimulus primary elections are usually the domain of highly engaged core voters (Campbell 1960; Göbel 2020). The results show that interpersonal communication reflects this notion of a concentric electorate.

We are not aware of any prior research that studies signals in communication with respect to electoral participation. However, related research finds that individuals' web browsing records are not more informative about turnout than random guessing (Bach et al. 2020). Other research that investigates signals in communication with respect to arguably more salient social identities, such as gender, report similar accuracy values of 63% (Beltran et al. 2020).

While the metrics of our model might seem small for turnout, especially compared to party preference, they are nontrivial in light of related work. Albeit only moderately, communication reliably conveys information about turnout in low-stimulus elections and by implication about highly engaged citizens who are most likely to socially identify as voters.

4.5.1 Additional Analyses

In this subsection, we consider consequences of choices we made regarding the training and evaluation of our neural network model. This discussion informs the robustness of our findings, the additional potential of our approach, and the nature of information that is relevant for predictive performance.

Our main results rest on an evaluation of classification performance on balanced data. Figure 29 in the appendix replicates the evaluation step over all models using the original, imbalanced test folds. Since we are now dealing with class imbalances, our attention should be on the PR-AUC (Cranmer & Desmarais 2017). This metric remains high for party preference and primary election turnout. For general election turnout it decreases further, substantiating our finding that interpersonal communication is not informative in this regard. Thus, we conclude that the decision to balance test folds is inconsequential for our results.

To maximize the available communication we decided to pool all text messages over time. This disregards time discrepancies in the recording of messages and associated outcomes. As a result, our findings might be blurred by intermittent changes in partisan attachment or the inclusion of post-election communication. Figure 30 in the appendix presents results using various temporal subsets for training and evaluation. For party preference, metrics remain mostly within the confines of the uncertainty intervals of our main results, suggesting that communicative signals

of voter preferences are quite stable over time. We obtain similar results for primary election turnout, even with a one-year lead. This highlights that communicative signals of turnout may prove useful in social media-based political forecasting.

We also decided to pool data geographically across states. Figure 31 in the appendix presents results from models alternately trained on data from all states, except one, which was held out for evaluation. While metrics degrade only slights for party preference, they do so substantially with regard to turnout. This reveals a distinctly local component in interpersonal communication as regards electoral characteristics.

Our focus in this paper is communication-specific. Accordingly, we have mapped person-level outcomes to corresponding text sequences. In Figure 32 in the appendix we explore what happens when we project resulting text predictions back at the person-level. Simultaneously, we examine the effect of restricting our test sample to individuals with increasing amounts of text sequences. Averaging out-of-sample predictions from our main models at the person-level substantially increases classification accuracy for all our outcomes. The amount of available communication per person, however, only affects party preference predictions. These results highlight the potential of interpersonal communication for measuring citizens' electoral characteristics.

Finally, we assess the importance of distinctly political as opposed to general communication for signaling voter preferences in participation. For these purposes, Figure 33 in the appendix alternately trains and evaluates the neural network on all and strictly political text sequences. For party preference, this exercise reveals that both political and non-political communication carry important signals. For turnout, the analysis points to the distinction between political and non-political as well as differences within political communication as important signals.

4.5.2 Qualitative Insights

We now present insights on the actual communication content that signals voter preferences and participation. Given the complexity of neural networks this is not easily extracted from the trained models. After all, whether a specific word is predictive of an outcome depends on its precise context, position, and semantic relationship with other words in a text sequence. Local Interpretable Model-agnostic Explanations offer a way to gain a qualitative understanding of signals captured by such complex models (Ribeiro et al. 2016b).

This procedure uses the trained network to predict the outcome of permutations of one text sequence. Features in these permutations are weighted by their proximity to the original text. A simple model is then applied to these weights and associated predictions to obtain a score that represents each word's contribution

to the predicted outcome. This provides only a local understanding of signals for specific communication examples, though. To understand which words are globally predictive of an outcome class, we apply this procedure to a random selection of 100 text sequences, sum the contribution scores of distinct words across texts (Ribeiro et al. 2016b), rank words accordingly, and extract the top 30. We repeat this for 100 resamples of text sequences, yielding 100×30 rankings. Finally, we employ a rank aggregation algorithm to extract from this the top globally predictive communicative features. This is done for every outcome, using a model close to average metrics and drawing text sequences from the test fold.

Table 4: Top predictive communicative features.

Outcome	Class	Type	Features
Party preference	Republican	Word	dems, msm, agenda, tds, democrats, conservative, socialist, great, potus, leftist, illegals, liberals, media, praying, hillary, patriots
		Hashtag	#god, #trump2020, #walkaway, #smallbusiness, #tcot, #maga, #jesus, #foxnews, #winning, #livepd, #fakenews, #2a ,
		Mention	@dailycaller, @breitbartnews, @realjameswoods, @mattwalshblog, @atensnut, @sebgorka, @greggutfeld, @realcandaceo
	Democrat	Word	trump's, criminal, republican, bernie, donald, mitch, democracy, corporate, republicans, fucking, donny, lies, russian, ugh, wh
		Hashtag	#notmypresident, #resist, #bernie2020, #notmeus, #bluewave, #moscowmitch, #jobs, #trumpisamoron, #lgbtq, #climatechange,
		Mention	@huffpostpol, @peterdaou, @maddow, @tulsigabbard, @msnbc, @davidmweissman, @ddale8, @petebuttigieg, @therickwilson
General election turnout	Not voted	Word	diabolical, premed, tagovailoa, hungover, decentralized, ticktock, shotguns, nietzsche, paternity, blockchain, nightcore, wod
		Hashtag	#vaccineswork, #quotestoliveby, #fauxnews, #namaste, #freewill, #destiny2, #pregnancy, #freepalestine, #stopthemadness
		Mention	@nireyal, @nhlflyers, @yrysbryd, @abigaildisney, @mikeguilday @claicyuzolt, @timinhonolulu, @tomheartstanks, @karoli
	Voted	Word	communists, intermediate, geesh, mishaps, obscenity, antilgbtq, goverment, santorum, ransomware, heretic, roster, trmp, pinellas
		Hashtag	#philanthropy, #dems, #iamantifa, #potus45, #closethecamps, #enemyofthepeople, #2a, #republicans, #constitution #gratitude,
		Mention	@carlosgsmith, @fernandoamandi, @laurpicks, @autumnsgrammy, @truthout, @lynnv378, @psychdr100, @omarbula, @kstreet111,
Primary election turnout	Not voted	Word	though, idk, myself, af, douche, dollars, salad, tesla, life, bitches, gym, honestly, bro, themselves, democrat, miss, rump, agreed
		Hashtag	#spirituality, #free, #starbucks, #trumprussia, #thebachelorette, #womeninstem, #blacklivesmatter, #bored, #cousins, #jesus
		Mention	@floridabbhour, @themeparkbeer, @ericmmatheny, @papabirdjake, @efawcett7, @ign, @citrix, @dolphinstalk, @nytopinion, @thescore
	Voted	Word	trumputin, party's, insurance, florida, liberal, conman, tallahassee, crooked, nazi, cult, terrorist, medicare, nra, illegals, gillum, fund
		Hashtag	#votered, #beavoter, #draintheswamp, #kag, #45, #buildthewall, #bluewave2018, #bringithome, #worstpresidentever, #vote
		Mention	@fladems, @mediaite, @dearauntcrabby, @nbcnews, @marcorubio, @maydaymindy9, @rondesantisfl, @dearauntcrabby, @gop

Table 4 presents the top predictive words, hashtags, and mentions of each class for every outcome. In line with our theoretical background, features predictive of party preference reflect individuals' social identification with partisan groups. For

both parties, references to the opposing side, accompanied by derogatory terms, are instructive about attachment. This highlights comparisons and a focus on differences between groups. Shared conceptions and concerns, such as “conservative”, “#god”, “#2a” (the right to keep and bear arms), or “#smallbusiness” for Republicans and “democracy”, “#resist”, “#climatechange”, “#lgbtq” for Democrats are the subject of talk and similarly telling about party prefernces. Proponents of both parties are also recognizable by (more positive) references to their own group, such as corresponding political elites or media outlets.

The top predictive features are much more ambiguous for general election turnout. This mirrors the poor classification performance and shows difficulties of the model to identify signals that are intuitively telling about voter participation in salient elections. Political features are somewhat more pronounced among terms predictive of voting. However, since political communication occurs also among non-voters (Göbel 2020), this by itself is not very instructive.

For primary election turnout, the divide between non-political and political terms seems much more informative about voter participation. The important difference here is that political features align strongly with partisan identification. Features such as “party’s”, “medicare”, “nra”, “liberal”, “ìllegals”, “#buildthewall”, “cult/#45”, (a pejorative term for Trump supporters), “trumputin”, or “#worstpresidentever” and references to group prototypes all reflect communicative patterns that we have seen for party preference. This reflects theoretical considerations that behavioral consequences of partisan identification inform voter participation. That this is only the case for primary election turnout is plausible, since strong partisans are also more likely to be highly engaged voters (Huddy et al. 2015).

In addition, signals of participation also support the view that being a voter functions as a political identity for this group of core voters. Features such as “#beavoter” or “#vote”, that are strongly predictive of voting, are particularly consistent with this. Further into the list of instructive features about voting we find many more related terms such as “#civics”, “donating”, “#encouragement”, “#whenweallvote”, “#forthepeople”, or “#everyvotecounts”.

These terms do not occur as signals of abstention. Instead, use of the first- and third-person pronouns “myself” and “themselves” is emphasized, which highlight the individual and differentiation from others (Slapin & Kirkland 2020). The features “idk” and “#bored” further reflect uncertainty and alienation.

Specific communication examples are offered in Figure 17. The colored texts illustrate the predictive contribution of words in context as regards party preference and primary election turnout. The signals highlighted in these examples further substantiate our findings.

Figure 17: Examples of model explanations (colors indicate support for the predicted class).

Table 3: Examples of model explanations (colors indicate support for the predicted class).

Outcome	Actual class	Predicted class	Example
Party preference	Republican	Republican	#babieslivesmatter i am all for women's choice but this is murder terminating a pregnancy should be illegal
	Republican	Republican	schools are already gun free and it is not working having armed teachers seems very reasonable
	Democrat	Democrat	because military style weapons are being used in mass shootings ems personnel prepare for war type wounds
	Democrat	Democrat	@webstercongress in your district women make 88 of a man's dollar need paycheck fairness act now #ru4fairpay
	Republican	Democrat	@realdonaldtrump holds the record for longest shutdown in history maybe congress can finally work together
	Democrat	Republican	obama deported illegal immigrants because that is what the government is supposed to do impose the law
Primary election turnout	Not voted	Not voted	american justice lost its meaning your partisan blocking collusion and greed has caused this
	Not voted	Not voted	biden's morals bother me and the other rapist will be far worse it is an absolutely shit choice
	Voted	Voted	it is not ideal to be voting during #coronavirus but if you are healthy embrace your civic duty
	Voted	Voted	by voting for this patriot you will #voteredtosaveamerica
	Not voted	Voted	great more people who get free medical care but do not contribute to our society
	Voted	Not voted	she does not realize how her personality disconnects voters

Contradicting — Supporting

Note: Examples have been slightly paraphrased to prevent identification of individuals in our data.

4.6 Conclusion

Our empirical analysis suggests that interpersonal communication conveys substantive information about voter preferences and participation. US citizens' talk is profoundly informative about their party preference. It is also insightful about whether citizens participate at low-stimulus elections, such as congressional primaries, and thus about high voter engagement. We demonstrate that signals in communication allow to infer party preference with 72% and voting at a primary election with 60% accuracy.

To be sure, our empirical evidence does not speak to causal effects of interper-

sonal communication on electoral characteristics. Similarly, it does not offer evidence on forecasting vote choice and turnout using communication. The evidence we present in this paper is exploratory and descriptive. That communication content transmits discussant characteristics acts as an important explanation for political effects of interpersonal communication. By showing that communication signals voter preferences and participation we inform the plausibility of this mechanism.

Our findings also speak to language use in politics. The communicative signals that reveal voter preferences and participation pertain to polarized political views, shared concerns, and common behavioral expectations that delineate political groupings. While this is especially the case for partisan identity, it is to some extent also apparent as regards identification as a voter. This suggests that political identities fundamentally structure citizens' talk.

Our study demonstrates the utility of artificial neural networks for studying political dimensions of citizen communication. We believe that our findings are particularly promising as regards the future development of communication-based measures of voter preferences and participation. While the classification performance of our model leaves much room for improvement at this stage, we hint at potential enhancement by projecting communication-specific predictions back at the person-level. Further research in this direction will prove useful for identifying likely voters on Twitter. Relevant information is typically unknown on this platform and may help improve social media-based election polling.

Finally, we want to raise awareness of some limitations of our study that may find consideration in future research. First, our inferences are based on communication on Twitter. However, Twitter is only one among several social networking platforms and results may differ for others. More importantly, we do not know to what extent the content of talk on Twitter generalizes to real-life conversations and whether our findings apply there as well. Second, data requirements force us to exclude unregistered voters from our study, of whom many certainly have party preferences and are not chronic nonvoters. Registered voters and potentially also Twitter users may be more involved in political conversation than the electorate at large. To the extent that political communication drives classification performance, our findings, though focusing exclusively on citizens, may still prove optimistic. Lastly, we focus on the US where political identities are highly developed. Our finding that predictive signals rest on political identities begs the question whether interpersonal communication conveys electoral characteristics in contexts where such identities are less developed.

A Appendix Section 2

A.1 Baseline Model

The baseline model uses a bag-of-words (BOW) text representation. Each word or token is counted in its own column and then bound together in a document term matrix (DTM). On top, a simple neural network is trained with a single neuron output layer and a sigmoid activation function, which is technically equivalent to logistic regression but much more data efficient. The model formula goes as follows:

$$f(x) = \frac{1}{1 + e^{-\theta x}}$$
$$p(y = 1, x) = f(x)$$

The output is often thought to be a probability score, but it is more realistically a discrimination score, with new samples being simply interpolated. As a result, unseen data with unknown distributions makes it difficult for any model to generalize well and allow reasonable extrapolations or generalization.

During training, input vectors are propagated layer by layer through the network until they meet the output layer. Using a loss (objective, cost) function, the network's predictions are rewarded for correctly defined data and penalized for misclassified data [@goldberg2]. An example of cost function for the regression task is a root mean squared loss (L_2):

$$\ell(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x) - y)^2$$

For a binary outcomes crossentropy is the typical choice of loss function, but this can be easily changed including multiclass and multitask predictions.

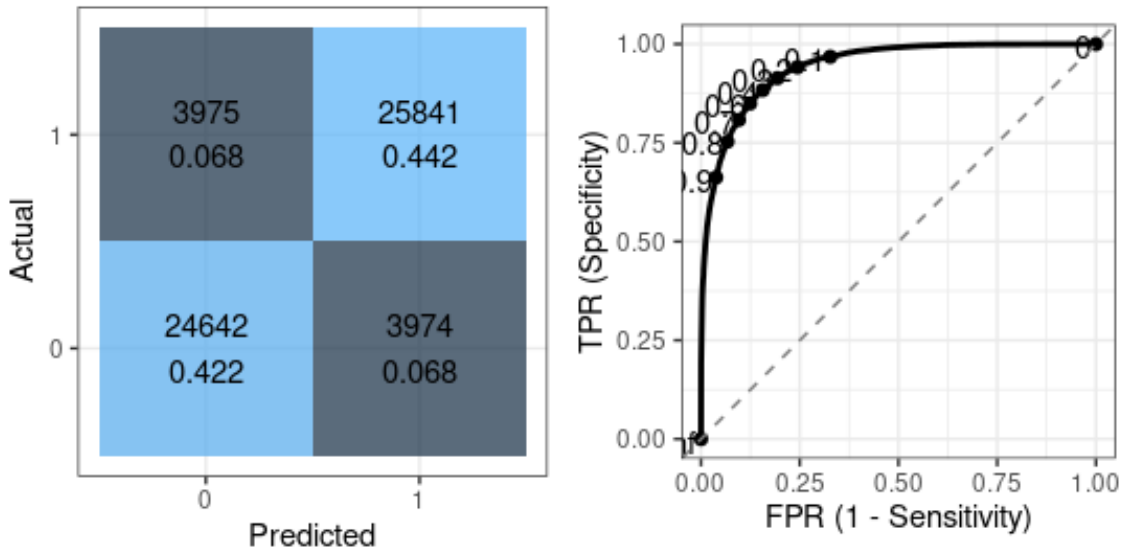
$$\ell(f) = -(y \log f(x) + (1 - y) log(1 - f(x)))$$

During the training process, the weights in a neural network are modified using techniques such as gradient descent, which finds the minimum of a function by constructing partial derivatives to update each parameter of a neural network. This procedure is repeated iteratively over a number of mini-batches and epochs.

B Appendix Section 3

B.1 Text Classification Model for Party Affiliation

Figure 18: Ideology text classifier test performance (out of sample)



B.2 Causal Estimates for Filter Bubbles

Table 5: ATE Linear Model (Pooled)

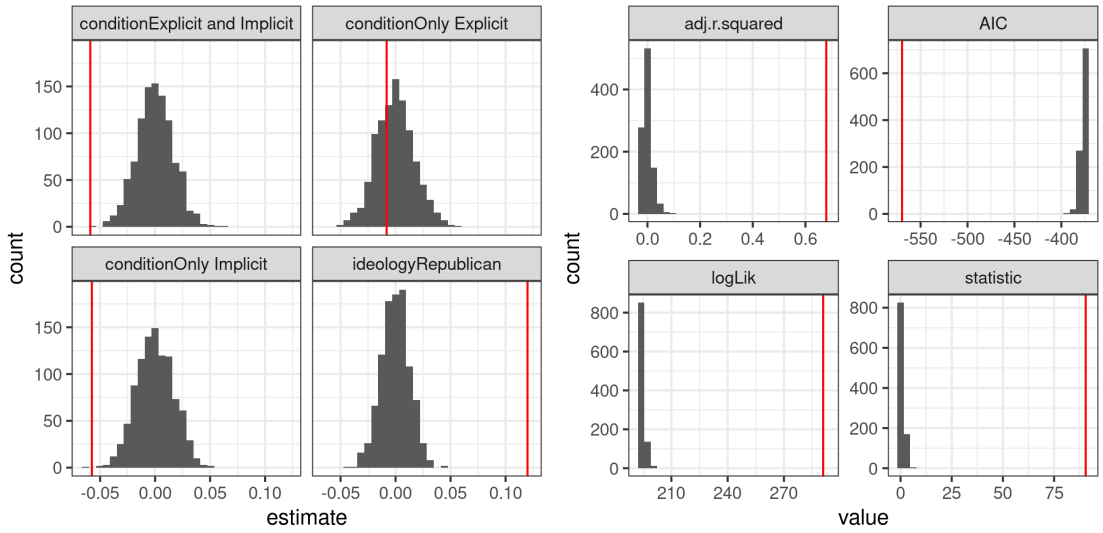
	Dependent variable:		
	Percent Crosscutting Content		
	(1)	(2)	(3)
Only Explicit	−0.004		−0.008
	(0.016)		(0.010)
Only Implicit	−0.059***		−0.057***
	(0.017)		(0.010)
Explicit and Implicit	−0.049***		−0.059***
	(0.016)		(0.009)
Conservative		0.118***	0.120***
		(0.008)	(0.007)
Constant	−0.001	−0.095***	−0.065***
	(0.012)	(0.006)	(0.008)
Observations	170	170	170
R^2	0.110	0.566	0.687
Adjusted R^2	0.094	0.563	0.679
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 6: CATE Linear Model (Liberal vs Conservative)

	Dependent variable:	
	Percent Crosscutting Content	
	(Liberal)	(Conservative)
Only Explicit	−0.023**	0.004
	(0.011)	(0.015)
Only Implicit	−0.055***	−0.060***
	(0.011)	(0.016)
Explicit and Implicit	−0.084***	−0.041***
	(0.011)	(0.014)
Constant	−0.056***	0.047***
	(0.008)	(0.011)
Observations	74	96
R^2	0.492	0.222
Adjusted R^2	0.470	0.196
Note:		*p<0.1; **p<0.05; ***p<0.01

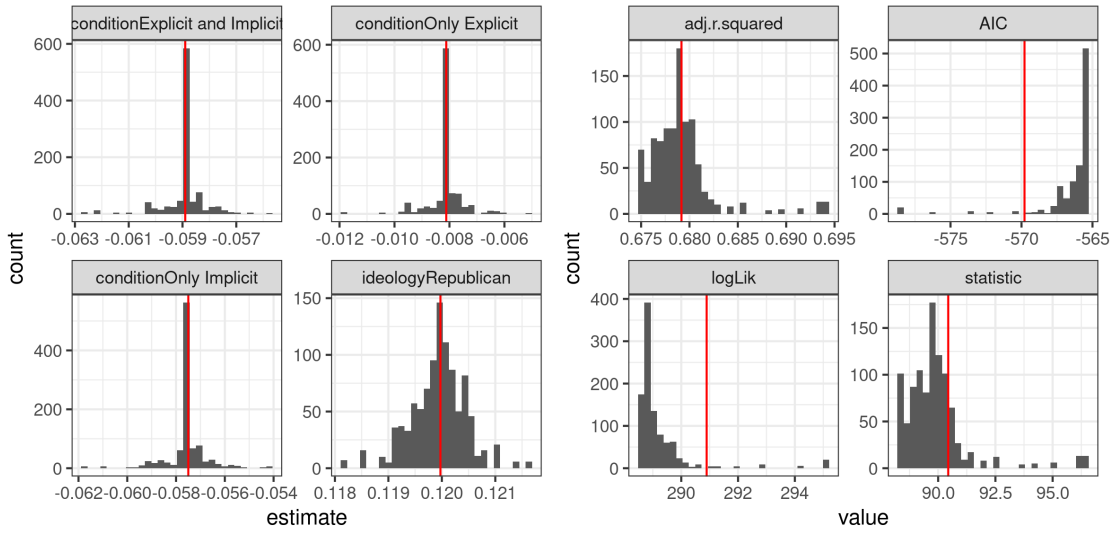
B.3 Robustness and Validation of Estimates

Figure 19: Randomized inference



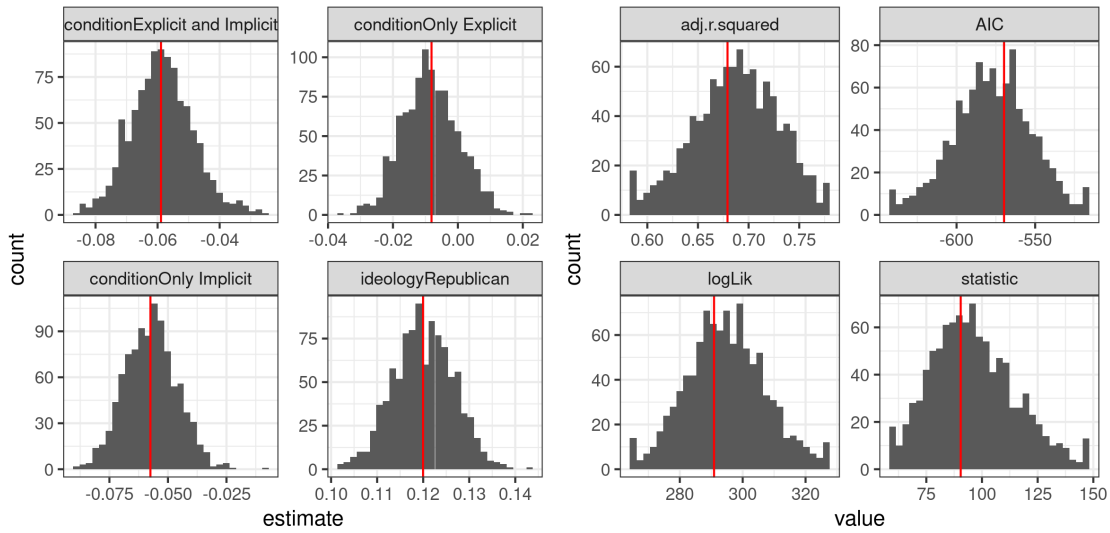
Note: We compare our model in red with the distributions of 10000 random systems that realize around zero and show that the data behind our models is far away from a random observation.

Figure 20: Leave-one-out validation



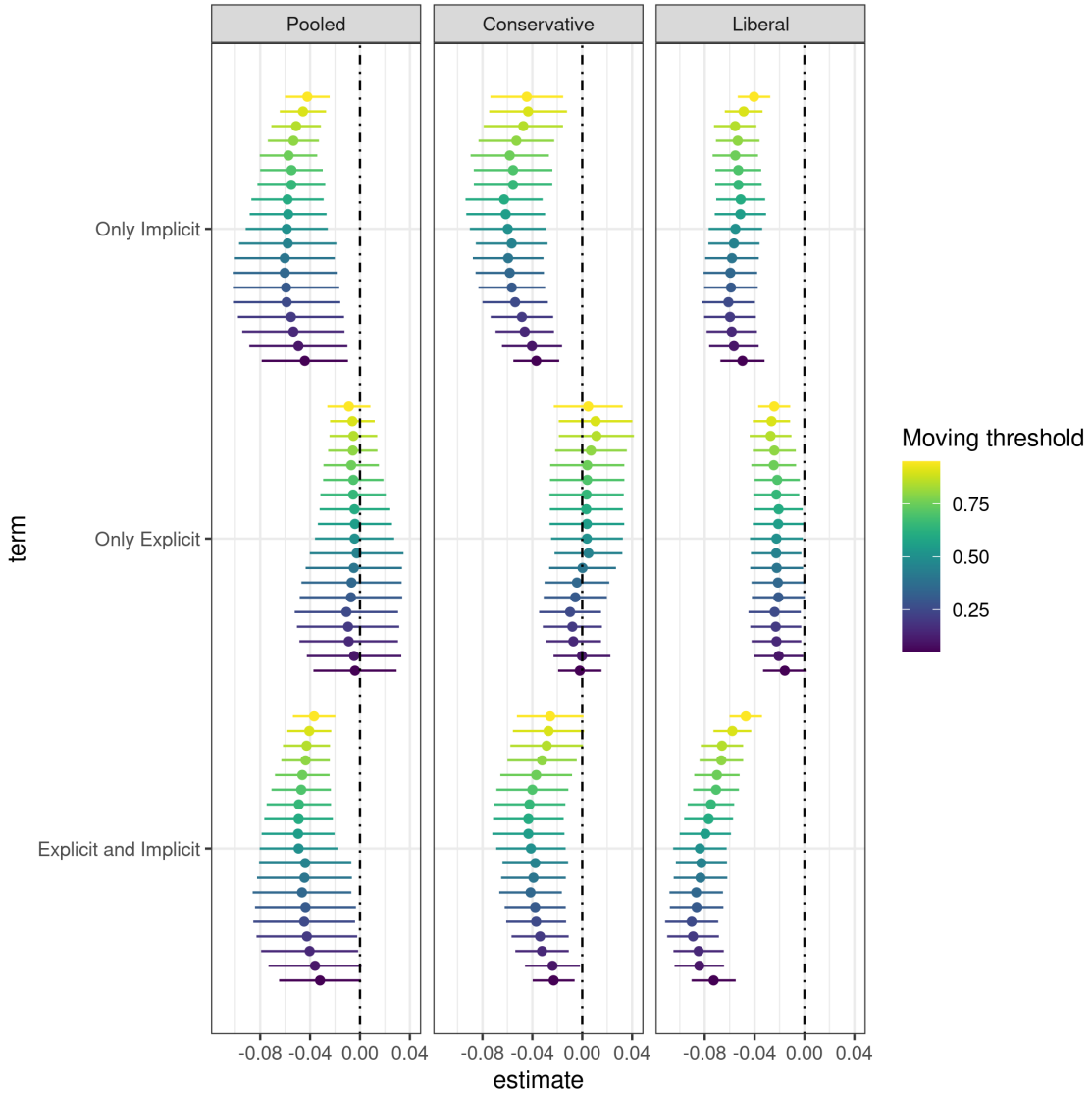
Note: We measure the impact of single observation on the stability of estimated parameters. Although, we can observe little variation in the parameter estimates, a single bot is unlikely to bias the result substantially given a simulation of 10000 trials.

Figure 21: Bootstrap with replacement



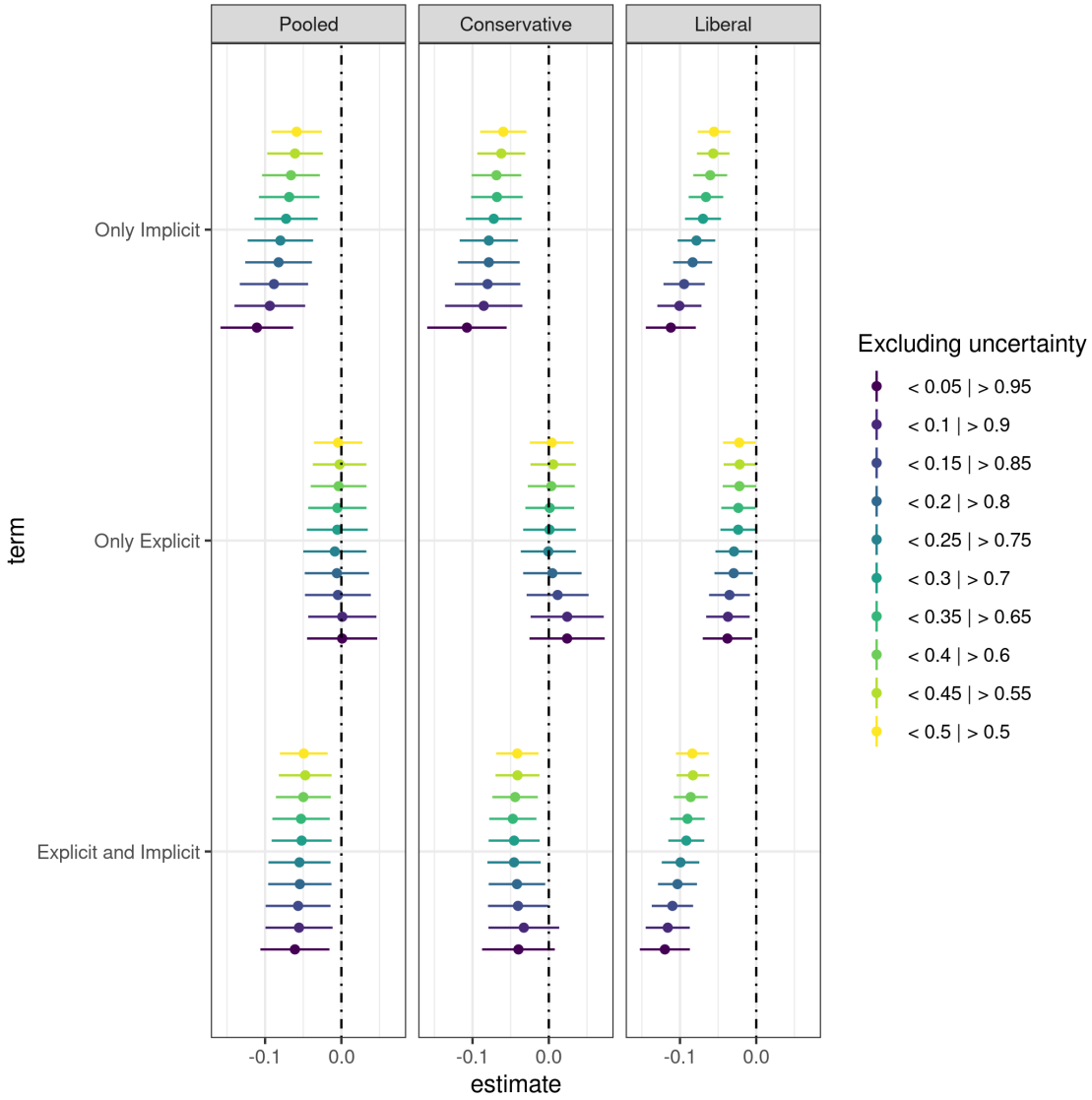
Note: We use the bootstrap for estimating confidence intervals for complex statistical models or arbitrary functions whose variance properties are difficult to analytically derive. The provided boundaries from 10000 trials match the standard errors from a linear model. In addition we can check whether a underlying data sample is only one lucky realization of the unknown population, which seems not to be the case.

Figure 22: Moving threshold



Note: We wanted to clarify the impact of the probability threshold for our ideology measurement. The plot shows the average treatment for each condition in our experiment with varying probability threshold. By 5% steps we can observe only some minor variation the estimates of interest and suggest that picking a threshold of .5 is also adequate for this data with only minor impact of calibrating model predictions post-hoc.

Figure 23: Measurement uncertainty



Note: We wanted to clarify the impact of prediction uncertainty for our ideology measurement. The plot shows the average treatment for each condition in our experiment with varying amount of observations being excluded due to falling into the uncertainty boundaries. By 5% steps we can observe only some minor variation the estimates of interest and suggest that model uncertainty is not producing statistical artefacts that would require calibrating the predictions post-hoc.

C Appendix Section 4

C.1 Details on Data Collection

Our empirical analysis rests on US voter registration records that were combined with Twitter accounts. We focus on voter records from 6 US states: Arkansas, Colorado, Connecticut, Florida, Rhode Island, South Dakota. This selection is due to requirements posed by our approach for integrating voter records with Twitter data.

The voter records we rely on in this paper can be obtained directly from the Division of Elections of the state departments for a processing charge or downloaded online from a public collection of lists maintained by Tom Alciere, a former Republican state representative who was running for United States Senator in New Hampshire as a write-in in 2020.[12]. For South Dakota, we obtained a free copy for the 9th state legislative district from https://nationbuilder.com/voterfile.

Data originally compiled by goebel2020 provides the bulk of information for our analysis. These data include voter records from the state of Florida as of October 2017. Self-reported email addresses of registered voters are made publicly available in Florida's voter records. These email addresses were used to match individual records to Twitter users who allowed others to find them by their email address. At that time, email discoverability was the default on Twitter. goebel2020 programmed an automated matching routine that implemented this procedure via Twitter's contact synchronization feature. See the respective paper's Appendix for a detailed description on this method. In total, of 681,096 registered voters who reported an email 90,832 could be associated with a public Twitter account. As described in the paper, we limit our analyses to those who declared a party registration, were eligible to vote in the 2018 general and primary election and posted on Twitter (See Table 3 in the paper).

To further extend the data we compiled voter records from other states where personal identifiers, such as email addresses or phone number, are publicly available in the voter records. In this regard we identified voter records from Arkansas, Colorado, Connecticut, Rhode Island, and South Dakota, which mostly make landline telephone numbers of registered voters publicly available. To associate corresponding Twitter accounts, we relied on services by the identity resolution company FullContact, which also allows to find persons' Twitter accounts based on phone numbers. FullContact maintains centralizes person profiles for over 275 million US adults, including various contact and social media identifiers. Given the cost associated with this service and a tight budget, we could only obtain a limited amount

[12]See https://arkvoters.com/, https://coloradovoters.info/, https://connvoters.com/, http://flvoters.com/, https://rivoters.com/

of Twitter accounts associated with registered voters. In total, across states we processed contact information of 561,883 registered voters via the FullContact API, which returned associated Twitter accounts for 26,931 registered voters. Since landline numbers may be outdated in the voter records or at FullContact, we additionally checked for matching names and also removed protected and terminated Twitter accounts, leaving us with 16,444 accounts associated with registered voters. Again, we limit our analyses to those who declared a party registration, were eligible to vote in the 2018 general and primary election and posted on Twitter.

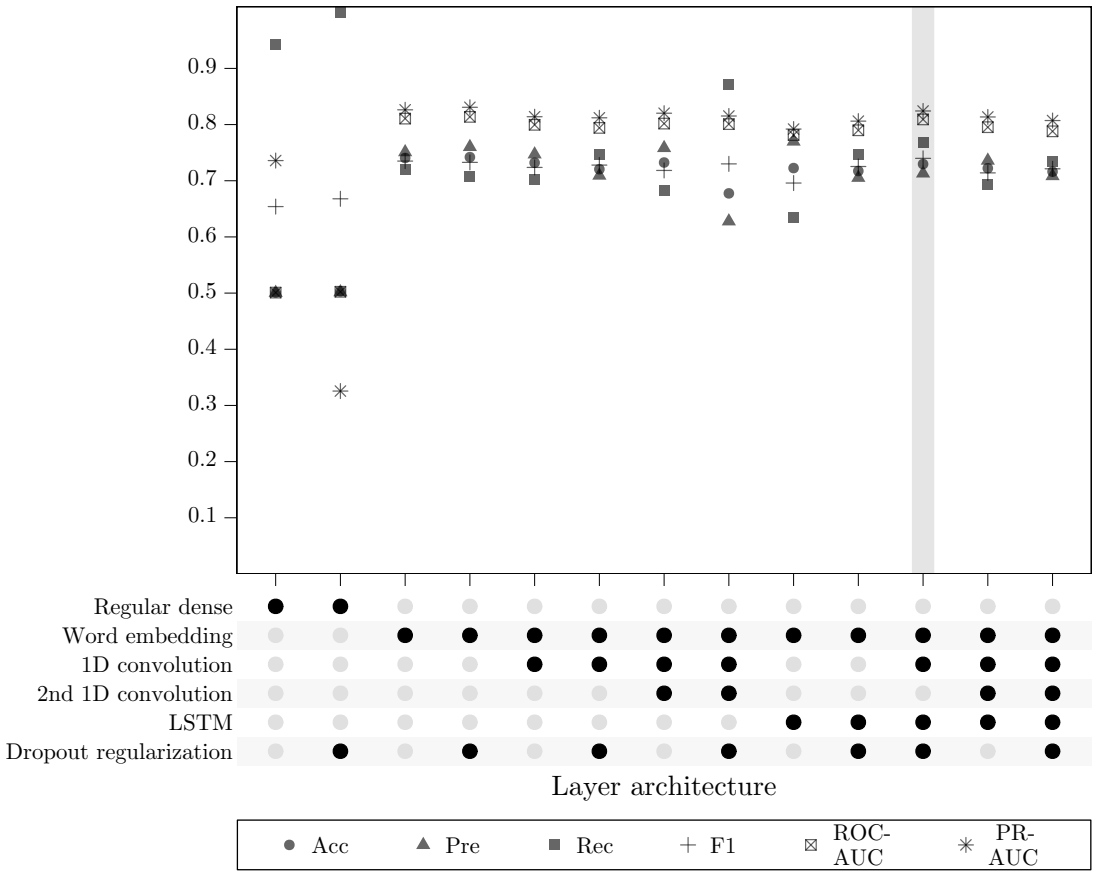
C.2 Comparison with Alternative Architectures

We assess the sensitivity of our results with regard to several layer architecture and hyperparameter choices. Since training and evaluating neural networks is a rather time consuming undertaking, we restrict assessments to the party preference outcome and do not use repeated cross-validation here.

Figure 24 pits our model (gray shade) against various combinations of layers, with and without regularization. All models were trained and evaluated using the same train and test split with a split ratio of 90/10. Hyperparameters were set as in our original model described above. Just as in the paper, the length of the text sequences was set to 300 and the model was processed in batches of size 256. 'Regular dense' describes a shallow model with just one simple hidden layer with 32 units. Except for this model, differences between various architecture choices are negligible. The shallow model fails utterly as it almost only predicts one class.

Batch size, which regulates the number of input texts processed by the model at once before updating weights and biases, is an important hyperparameter of neural

Figure 24: Results for multi layer architectures (party affiliation)

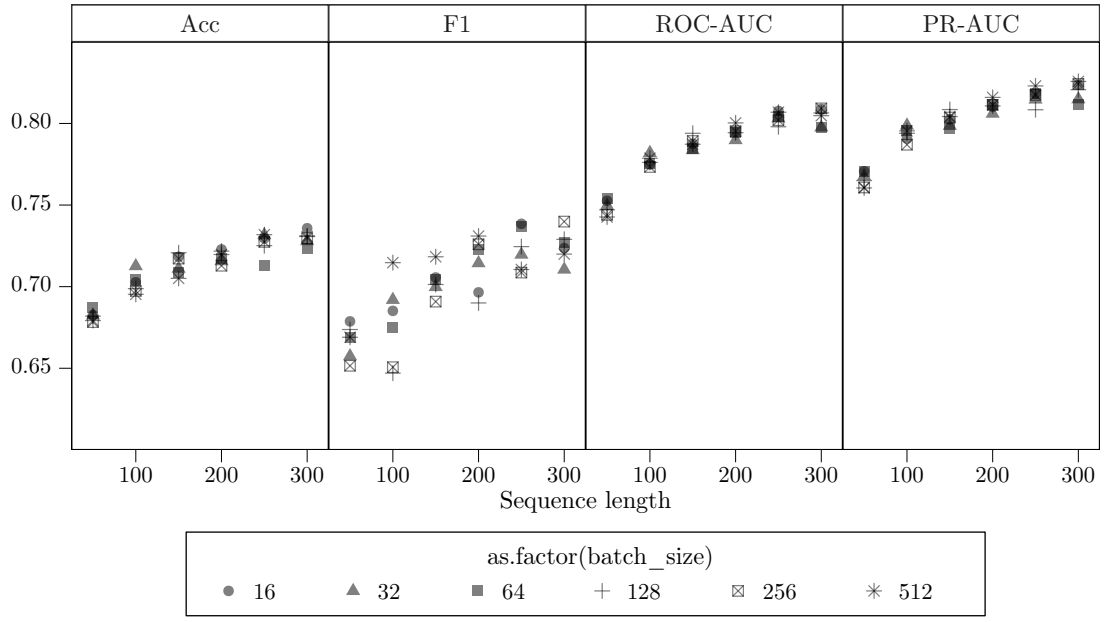


Note: Acc = Accuracy, Pre = Precision, Rec = Recall, ROC-AUC = receiver operator characteristic–area under the curve, PR-AUC = precision-recall-area under the curve.

networks. It governs the number of backpropagations and with that the results and their stability as well as runtime of the model. Larger batch sizes accelerate the training process but may increase the variance of results between runs of the same model. Given the amount of data, we picked a somewhat larger batch size of 256 to keep runtime manageable. To provide insights as regards the impact of that choice and its interplay with the length of the input sequence, we repeatedly ran our model with varying batch sizes over several sequence lengths. To maintain comparability between the model runs, we again use the same train and test split with a split ratio of 90/10. The results are shown in Figure 25. Differences between batch sizes in terms of the predictive performance of the model are minor and do not depend on the sequence length of the input.

Batch size dramatically impacts the runtime of the model, however, as is shown in Figure 26. In our case, training the neural network with a sequence length of 300 in under one hour is only possibly with a batch size of 128 or higher. Running our model with a batch size of 128 takes about 35 minutes whereas it takes about 26 minutes with a batch size of 256 and about 25 minutes with a batch size of 512. The difference between batch sizes 256 and 512 is negligible, the difference of 9 minutes per model run between batch size 128 and 256 is not. Recall that we ran our model 100 times for each outcome due to repeated cross validation. In total, this makes for a difference of 15 hours per outcome, and 45 hours over all outcomes. So while in

Figure 25: Predictive performance with different sequence lengths and batch sizes (party affiliation)



Note: Acc = Accuracy, Pre = Precision, Rec = Recall, ROC-AUC = receiver operator characteristic–area under the curve, PR-AUC = precision-recall-area under the curve.

principle we could have selected any batch size without repercussions on the results, only a larger batch size made it feasible to conduct repeated cross-validation within a bearable time horizon.

As concerns the hyperparameters pertaining to the central layers of the model, we repeatedly trained the model while changing core hyperparameters. Again, this was done using the same train and test split. When shifting a specific hyperparameter, we left the other values untouched as in the model described above. Figure 27 displays the hyperparameter shifts and their results. Hyperparameter values of our model in the paper are shaded in gray. Clearly, the choice of values for core hyperparameters in central layers of our neural network is inconsequential as regards predictive performance.

To go even further we repeated this exercise this time picking 25 random hyperparameter combinations within specified ranges. We let the following hyperparameters vary: the dimension of the embedding layer {25, 50, 100, 150, 200}, the filters {16, 32, 64, 128, 256}, kernel size {2, 3, 4, 5}, and strides {1, 2, 3, 4} of the convolution layer, the pool size {2, 3, 4, 5} and strides {1, 2, 3, 4} of the max pooling layer, and the units of the long short-term memory layer {16, 32, 64, 128, 256}. Figure 27 reassures that shifts to hyperparameters barely affect the overall results of our model. Notice, however, that there is still room for improving the recall metric through hyperparameter optimization, which we leave to future research.

Figure 26: Runtime with different sequence lengths and batch sizes (party affiliation)

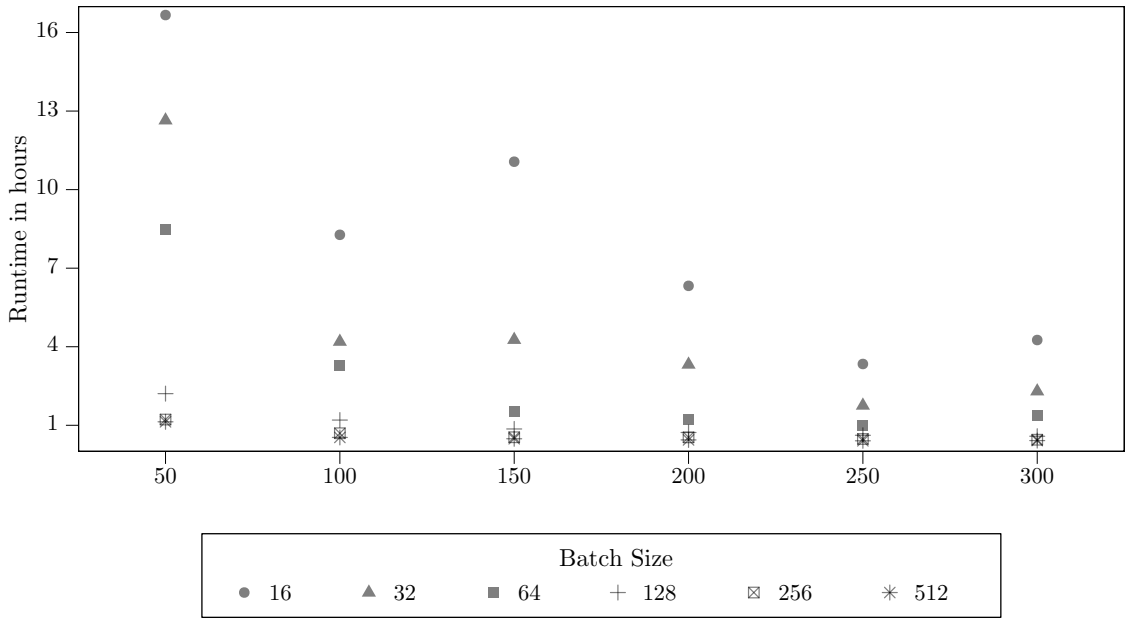
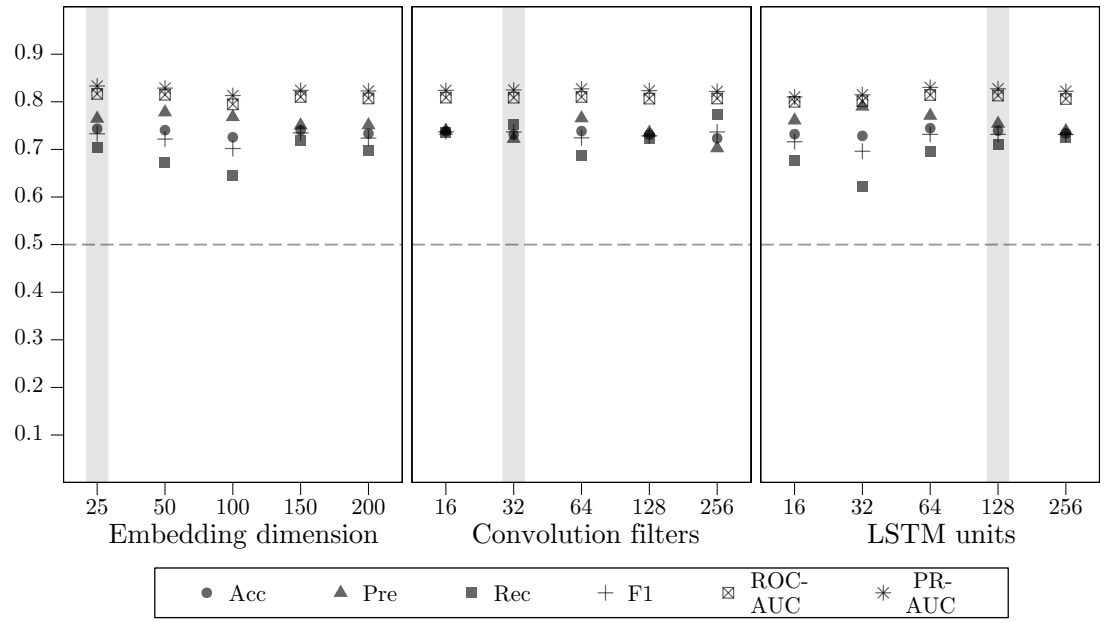
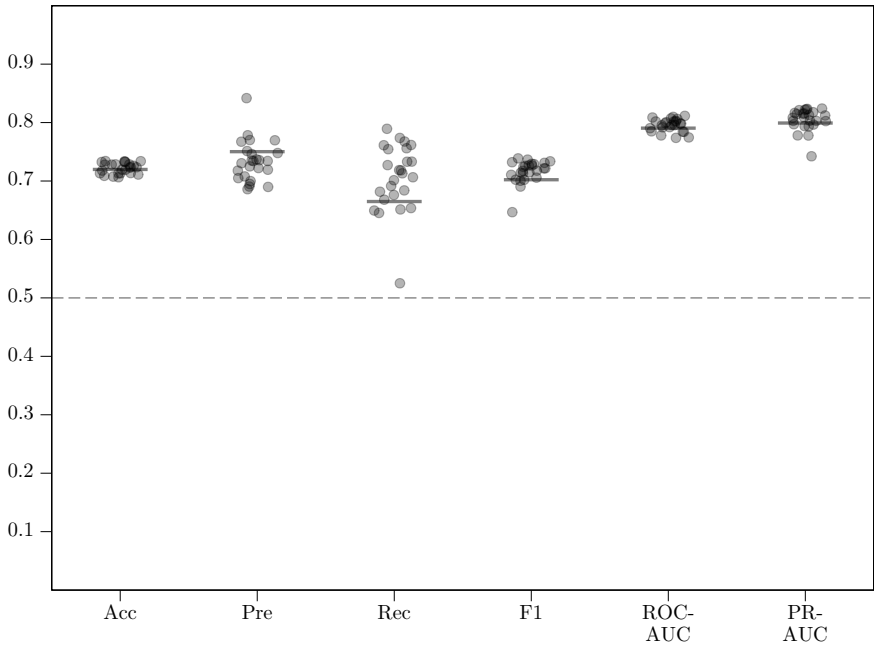


Figure 27: Fixed hyperparameter variation (party affiliation).



Note: Acc = Accuracy, Pre = Precision, Rec = Recall, ROC-AUC = receiver operator characteristic–area under the curve, PR-AUC = precision-recall-area under the curve.

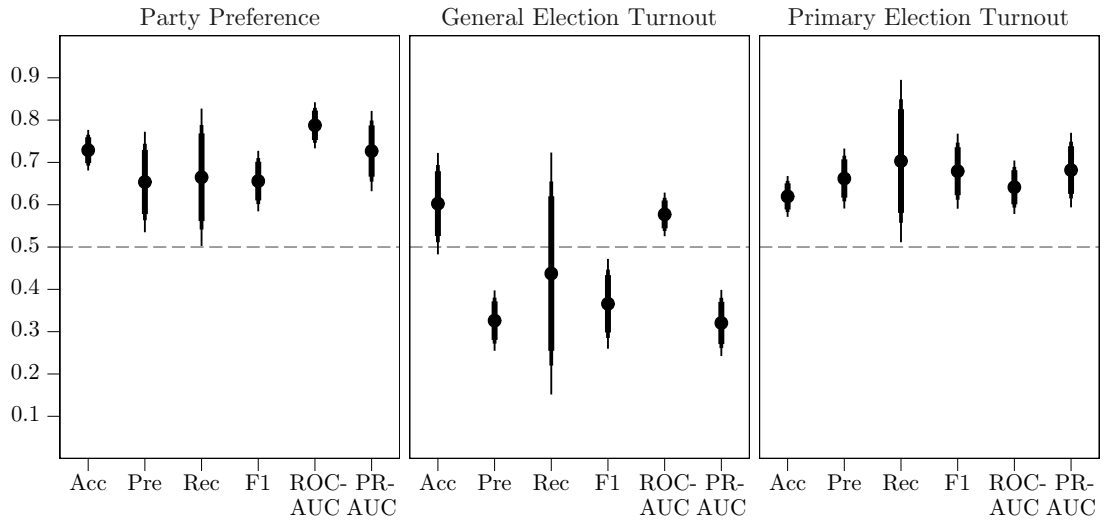
Figure 28: Random hyperparameter variation (party affiliation).



Note: Acc = Accuracy, Pre = Precision, Rec = Recall, ROC-AUC = receiver operator characteristic–area under the curve, PR-AUC = precision-recall-area under the curve. Results from our original model are indicated by the horizontal bar.

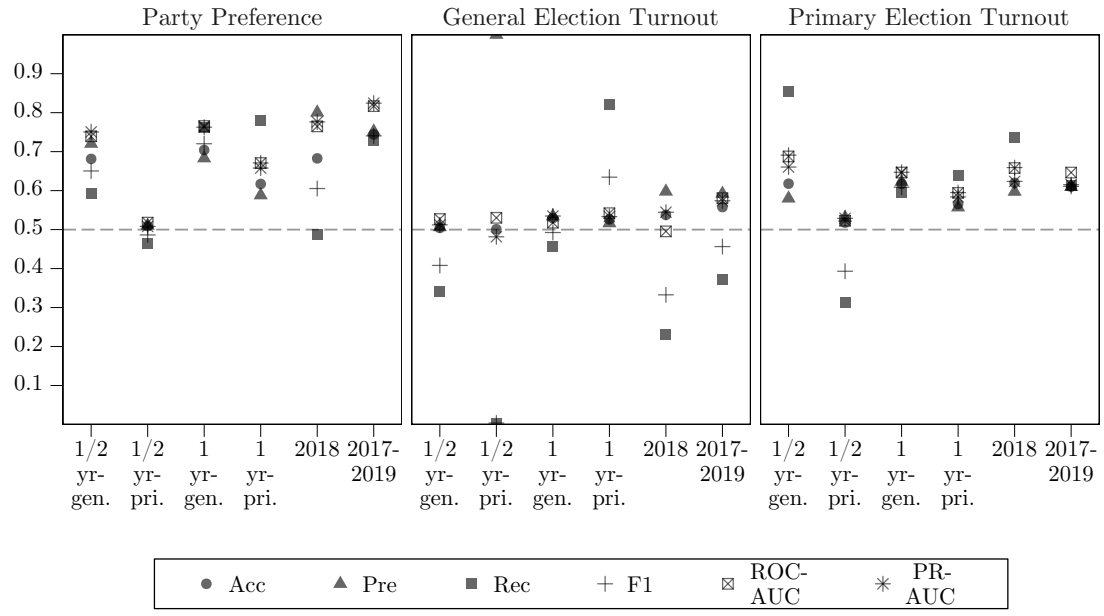
C.3 Additional Figures

Figure 29: Classification performance on unbalanced data.



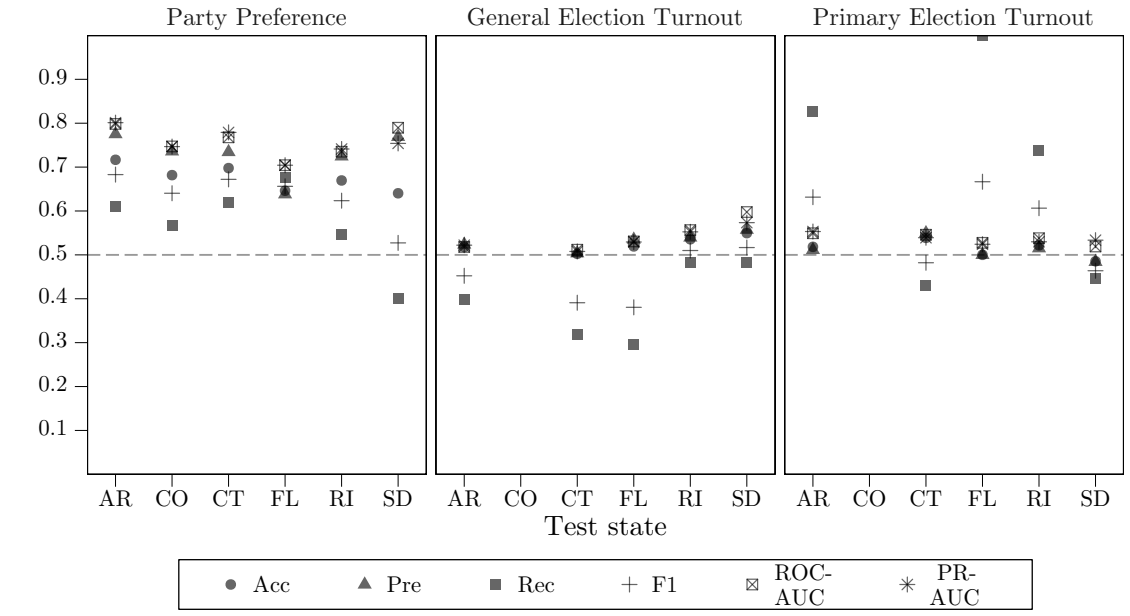
Note: Acc = Accuracy, Pre = Precision, Rec = Recall, ROC-AUC = receiver operator characteristic–area under the curve, PR-AUC = precision-recall-area under the curve. Mean and uncertainty intervals (90%, 95%, and 99%) are based on 10×10-fold cross-validation = 100 models for each outcome).

Figure 30: Classification performance across temporal subsets.



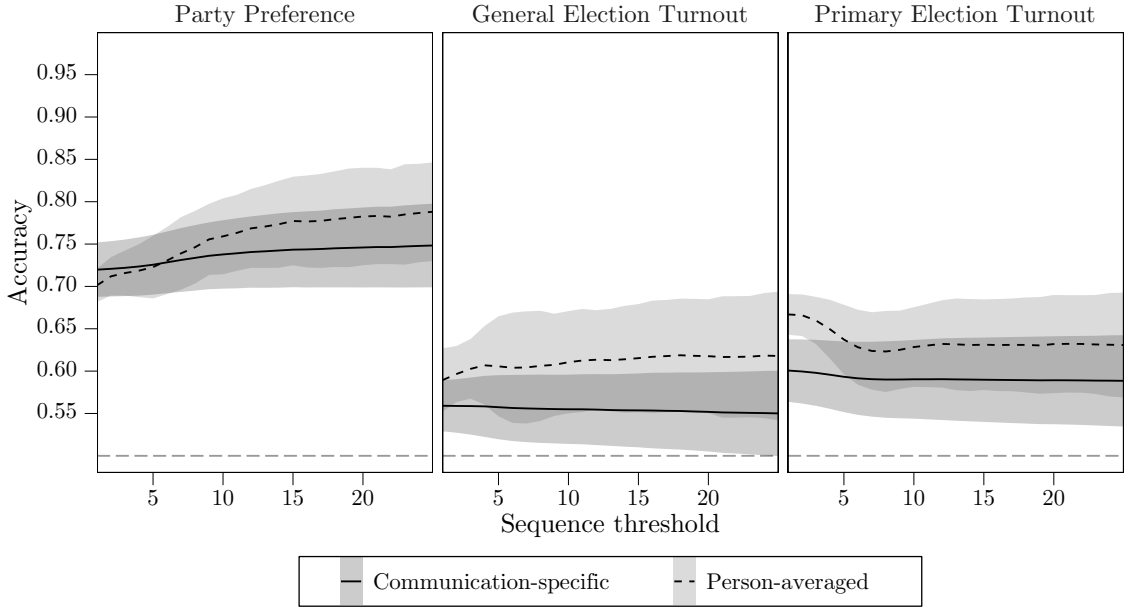
Note: Acc = Accuracy, Pre = Precision, Rec = Recall, ROC-AUC = receiver operator characteristic–area under the curve, PR-AUC = precision-recall-area under the curve.

Figure 31: Classification performance under varying geographical test data.



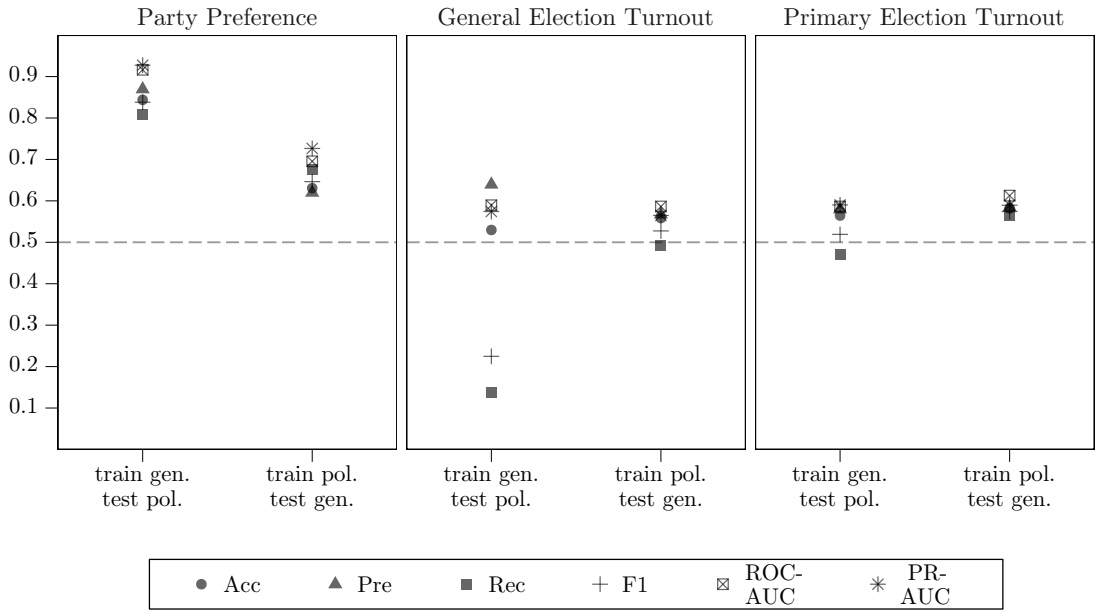
Note: Acc = Accuracy, Pre = Precision, Rec = Recall, ROC-AUC = receiver operator characteristic–area under the curve, PR-AUC = precision-recall-area under the curve.

Figure 32: Classification performance under varying activity thresholds and considering communication-specific and person-averaged predictions.



Note: Mean and uncertainty intervals (95%) are based on $10{\times}10 - fold cross - validation = 100 models for each outcome$).

Figure 33: Classification performance when alternating general and political train and test data



Note: Acc = Accuracy, Pre = Precision, Rec = Recall, ROC-AUC = receiver operator characteristic–area under the curve, PR-AUC = precision-recall-area under the curve. We rely on a dictionary-based classifier by goebel2020 to identify political text sequences.

Author Contributions

All of the publications mentioned above employ the same approach, which is text classification using neural networks. This is mirrored in the first article, which is exclusively committed to best practices in machine learning. Both collaborations in papers 2 and 3 depended significantly on this method and would have been impossible to accomplish without it.

The second paper was a two-year study in cooperation with Ben Guinaudeau and Fabio Votta. The burden could not have been completed by a single or even two people. Although it is impossible to list each accomplishments in detail, my own contributions may be summed up in three points. First, I had to build bots into our Twitter infrastructure, which took a long time. Second, I provided a cross-validated text prediction model that acted as a treatment policy for our social media bots. Third, I analyzed the data from our experiment, which includes various validity and sensitivity checks. Analysis of data was of course also supported by my colleagues. My duties revolved on machine learning and statistical inference models, and I was less engaged with the bot infrastructure, much to the delight of Ben Guinaudeau.

The third study was a collaboration with Sascha Göbel, who gathered sensitive data from Voter Registration files and their associated Twitter account communication. This nearly unbiased sample of data on communication and voting behavior in the real world enabled me to build machine learning models for prediction. Again, the contributions are many and difficult to isolate. The three primary elements that I contributed into the project were as follows: first, I built a model infrastructure to train and register models with various backends. Second, try alternative ways to prepossessing, batching, and modeling. Third, cross-validated classifiers for a variety of outcomes such as party affiliation and voter participation were developed. Despite the fact that Sascha was involved in every phase of the model development, he focused more on the data inputs as well as the insights on feature importance and relevance. In contrast to the second article, where writing was more equally distributed, Sascha contributed the most to the final paper in this capacity.

References

Aıvodji, U., Bidet, F., Gambs, S., Ngueveu, R. C., & Tapp, A. (2019). Adversarial training approach for local data debiasing. *arXiv preprint arXiv:1906.07858*.

Ali, M., Sapiezynski, P., Korolova, A., Mislove, A., & Rieke, A. (2019). Ad delivery algorithms: The hidden arbiters of political messaging. *arXiv preprint arXiv:1912.04255*.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Arlot, S., & Lerasle, M. (2016). Choice of v for v-fold cross-validation in least-squares density estimation. *J. Mach. Learn. Res.*, *17*(1), 7256–7305.

Ashish Bansal. (2019). Challenges in Recommender Systems at Twitter Scale. http://tiny.cc/ncmbtz

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2020). Predicting voting behavior using digital trace data. *Social Science Computer Review*, in press.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, *348*(6239), 1130–1132.

Balayn, A., & Gürses, S. (2018). Beyond debiasing - regulating ai and its inequalities. *Delft University of Technology, the Netherlands*.

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–34.

Bandy, J., & Diakopoulos, N. (2020). Auditing news curation systems: A case study examining algorithmic and editorial logic in apple news. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 36–47.

Banerjee, B. (2017). Machine learning models for political video advertisement classification.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, *23*(1), 76–91.

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, *29*(1), 19–42.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Baxter, J. (2000). A model of inductive bias learning. *Journal of artificial intelligence research*, *12*, 149–198.

Beam, M. A. (2014). Automating the news: How personalized news recommender system design choices impact news reception. *Communication Research*, *41*(8), 1019–1041.

Beauchamp, N. (2017). Predicting and interpolating state-level polls using Twitter textual data. *American Journal of Political Science*, *61*(2), 490–503.

Bechmann, A., & Nielbo, K. L. (2018). Are we exposed to the same "news" in the news feed? an empirical analysis of filter bubbles as information similarity for danish facebook users. *Digital journalism*, *6*(8), 990–1002.

Beck, P. A., Dalton, R. J., Greene, S., & Huckfeldt, R. (2002). The social calculus of voting. Interpersonal, media, and organizational influences on presidential choices. *American Political Science Review*, *96*(1), 57–73.

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854.

Beltran, J., Gallego, A., Huidobro, A., Romero, E., & Padró, L. (2020). Male and female politicians on Twitter. A machine learning approach. *European Journal of Political Research*, in press.

Bergmeir, C., & Benıtez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213.

Berk, R. (2012). *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

Beskow, D. M., Kumar, S., & Carley, K. M. (2020). The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing & Management*, *57*(2), 102170.

Bhatti, Y., Fieldhouse, E., & Hansen, K. M. (2020). It's a group thing. How voters go to the polls together. *Political Behavior*, *42*(1), 1–34.

Blais, A., & Achen, C. H. (2019). Civic duty and voter turnout. *Political Behavior*, *41*(2), 473–497.

Bolstad, J., & Dinas, E. (2017). A categorization theory of spatial voting. How the center divides the political space. *British Journal of Political Science*, *47*(4), 829–850.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*, 4349–4357.

Bonchi, F., Hajian, S., Mishra, B., & Ramazzotti, D. (2017). Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, *3*(1), 1–21.

Bono, M., & Melo-Pfeifer, S. (2011). Language negotiation in multilingual learning environments. *International Journal of Bilingualism*, *15*(3), 291–309.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Bozdag, E., & Van Den Hoven, J. (2015). Breaking the filter bubble: Democracy and design. *Ethics and information technology*, *17*(4), 249–265.

Breitung, J., & Knüppel, M. (2021). How far can we forecast? statistical tests of the predictive content. *Journal of Applied Econometrics*.

Bryan, C. J., Yeager, D. S., & O'Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences*, *116*(51), 25535–25545.

Bullock, J. G., Gerber, A. S., Hill, S. J., & Huber, G. A. (2013). *Partisan bias in factual beliefs about politics* (tech. rep.). National Bureau of Economic Research.

Burden, B. C., & Greene, S. (2000). Party attachments and state election laws. *Political Research Quarterly*, *53*(1), 63–76.

Calin, O. (2020). *Deep learning architectures. A mathematical approach*. Springer.

Campbell, A. (1960). Surge and decline. A study of electoral change. *Public Opinion Quarterly*, *24*(3), 397–418.

Carlson, T. N., Abrajano, M., & Bedolla, L. G. (2020). *Talking politics. Political discussion networks and the new American electorate*. Oxford University Press.

Castelvecchi, D. (2020). Is facial recognition too biased to be let loose? *Nature*, *587*(7834), 347–350.

Chakraborty, A., & Ganguly, N. (2018). Analyzing the news coverage of personalized newspapers. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 540–543.

Chang, C., & Masterson, M. (2020). Using word order in political text classification with long-short term memory models. *Political Analysis*, *28*(3), 395–411.

Chen, Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2020). Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240*.

Chen, Pacheco, Yang, & Menczer. (2021). Neutral bots probe political bias on social media. *Nature communications*, *12*(1), 1–10.

Chollet, F., & Allaire, J. J. (2013). *Deep learning with R*. Manning.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153–163.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Conference on Fairness, Accountability and Transparency*, 134–148.

Cichocka, A., Bilwicz, M., Jost, J. T., Marrouch, N., & Witkowska, M. (2016). On the grammar of politics. Or why conservatives prefer nouns. *Political Psychology*, *37*(6), 799–815.

Cinnamon, J. (2017). Social injustice in surveillance capitalism. *Surveillance & Society*, *15*(5), 609–625.

Cohen, R., & Ruths, D. (2013). Classifying political orientation on Twitter. It's not easy! *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 91–99.

Coos, A. (2018). Eu vs us: What are the differences between their data privacy laws? *endpointprotector.com*.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Cranmer, S. J., & Desmarais, B. A. (2017). What can we learn from predictive modeling? *Political Analysis*, *25*(2), 145–166.

Crawford, K., & Calo, R. (2016). There is a blind spot in ai research. *Nature News*, *538*(7625), 311.

Cunningham, S. J. (1995). Machine learning and statistics: A matter of perspective.

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

De Graaf, A., Hoeken, H., Sanders, J., & Beentjes, J. W. (2012). Identification as a mechanism of narrative persuasion. *Communication research*, *39*(6), 802–823.

Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2011). Language and ideology in congress. *British Journal of Political Science*, *42*(1), 31–55.

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms* (tech. rep.). Citeseer.

Dinas, E. (2014). Does choice bring loyalty? Electoral participation and the development of party identification. *American Journal of Political Science*, *58*(2), 449–465.

Druckman, J. N., Levendusky, M. S., & McLain, A. (2018). No need to watch: How the effects of partisan media can spread via interpersonal discussions. *American Journal of Political Science*, *62*(1), 99–112.

Egan, P. J. (2020). Identity as dependent variable. How Americans shift their identities to align with their politics. *American Journal of Political Science*, *64*(3), 699–716.

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). " i always assumed that i wasn't really that close to [her]" reasoning about invisible algorithms in news feeds. *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 153–162.

Eveland, W. P. J., & Hively, M. H. (2009). Political discussion frequency, network size, and "heterogeneity" of discussion as predictors of political knowledge and participation. *Journal of Communication*, *59*(2), 205–224.

Faddoul, M., Chaslot, G., & Farid, H. (2020). A longitudinal analysis of youtube's promotion of conspiracy videos. *arXiv preprint arXiv:2003.03318*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Feldstein, S. (2019). The road to digital unfreedom: How artificial intelligence is reshaping repression. *Journal of Democracy*, *30*(1), 40–52.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2019). Auto-sklearn: Efficient and robust automated machine learning. *Automated machine learning* (pp. 113–134). Springer, Cham.

Fieldhouse, E., & Cutts, D. (2020). Do as i say or do as i do? How social relationships shape the impact of descriptive and injunctive norms of voting. *British Journal of Political Science*, in press.

Fowler, J. H., & Kam, C. D. (2007). Beyond the self. Social identity, altruism, and political participation. *American Journal of Political Science*, *69*(3), 813–827.

Fricke, K., & Herzberg, P. (2019). Know your big data: De-biasing subsamples of large datasets for personality research using importance sampling and knn matching.

Friedman, J. (2014). Political epistemology.

Frimer, J. A., Skitka, L. J., & Motyl, M. (2017). Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *Journal of Experimental Social Psychology*, *72*, 1–12.

Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*.

Garbin, C., Zhu, X., & Marques, O. (2020). Dropout vs. batch normalization. An empirical study of their impact to deep learning. *Multimedia Tools and Applications*, *79*, 12777–12815.

Garrett. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, *14*(2), 265–285.

Garrett, Carnahan, & Lynch. (2013). A turn toward avoidance? selective exposure to online political information, 2004–2008. *Political Behavior*, *35*(1), 113–134.

Garrett, Gvirsman, Johnson, Tsfati, Neo, & Dal. (2014). Implications of pro- and counterattitudinal information exposure for affective polarization. *Human Communication Research*, *40*(3), 309–332. https://doi.org/10.1111/hcre.12028

Garrett, R. K., Gvirsman, S. D., Johnson, B. K., Tsfati, Y., Neo, R., & Dal, A. (2014). Implications of pro-and counterattitudinal information exposure for affective polarization. *Human communication research*, *40*(3), 309–332.

Garvie, C., & Frankle, J. (2016). Facial-recognition software might have a racial bias problem. *The Atlantic*, *7*.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gerber, A. S., Huber, G. A., & Fang, A. (2018). Do subtle linguistic interventions priming a social identity as a voter have outsized effects on voter turnout? Evidence from a new replication experiment. *Political Psychology*, *39*(4), 925–938.

Göbel, S. (2020). Voting and social media-based political participation. *Unpublished Manuscript. Available at https://osf.io/sjq4g/*.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345–420.

Gong, G. (1986). Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *Journal of the American Statistical Association*, *81*(393), 108–113.

Gottlieb, A. (2018). Private partisan, public moderate: Preference falsification on twitter. *Princeton University DataSpace*. https://github.com/alex-gottlieb/deepIdeology

Greene, S. (1999). Understanding party identification. A social identity approach. *Political Psychology*, *20*(2), 393–403.

Grimmer, Messing, & Westwood. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, *25*(4), 413–434.

Guess, A., Nyhan, B., & Reifler, J. (2018). Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council*, *9*.

Guess, Nyhan, Lyons, & Reifler. (2018). Avoiding the echo chamber about echo chambers. *Knight Foundation*, *2*.

Haim, M., Graefe, A., & Brosius, H.-B. (2018). Burst of the filter bubble? effects of personalization on the diversity of google news. *Digital journalism*, *6*(3), 330–343.

Hardin, R. (2009). Deliberative democracy. *Contemporary debates in political philosophy*, 229–246.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Hinds, J., & Joinson, A. N. (2018). What demographic attributes do our digital footprints reveal? A systematic review. *PLOS ONE*, *13*(11), 1–40.

Hogg, M. A., & Reid, S. A. (2006). Social identity, self-categorization, and the communication of group norms. *Communication Theory*, *16*(1), 7–30.

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, *2*(4), 100241.

Huckfeldt, R. (2005). Interdependence, density dependence, and networks in politics. *American Politics Research*, *37*(5), 921–950.

Huckfeldt, R., & Sprague, J. (1991). Discussant effects on vote choice. Intimacy, structure, and interdependence. *Journal of Politics*, *53*(1), 122–158.

Huddy, L. (2001). From social to political identity. A critical examination of social identity theory. *Political Psychology*, *22*(1), 127–156.

Huddy, L. (2013). From group identity to political cohesion and commitment. In L. Huddy, D. O. Sears, & J. S. Levy (Eds.), *The oxford handbook of political psychology* (pp. 738–773). Oxford University Press.

Huddy, L., Mason, L., & Aaroe, L. (2015). Expressive partisanship. Campaign involvement, political emotion, and partisan identity. *American Political Science Review*, *109*(1), 1–17.

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, *119*(1).

IJ. (2018). Statistics versus machine learning. *Nature methods*, *15*(4), 233.

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines. New evidence on group polarization. *American Journal of Political Science*, *59*(3), 690–707.

Johnson. (2010). The magic of groups identity: How predatory lenders use minorities to target communities of color. *Geo. J. on Poverty L. & Pol'y*, *17*, 165.

Johnson, Foster, D., & Stine, R. (2016). Impartial predictive modeling: Ensuring fairness in arbitrary models. *Statistical Science*, 1.

Johnson, & Fredrickson. (2005).

we all look the same to me" positive emotions eliminate the own-race bias in face recognition. *Psychological science*, *16*(11), 875–881.

Kalin, M., & Sambanis, N. (2018). How to think about social identity. *Annual Review of Political Science*, *21*, 239–257.

Kennedy, L. J., McCoy, P. A., & Bernstein, E. (2011). The consumer financial protection bureau: Financial regulation for the twenty-first century. *Cornell L. Rev.*, *97*, 1141.

Kenny, C. (1998). The behavioral consequences of political discussion. Another look at discussant effects on vote choice. *Journal of Politics*, *60*(1), 231–244.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*(2), 1137–1145.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805.

Kuhn, M., & Johnson. (2013). Data pre-processing. *Applied predictive modeling* (pp. 27–59). Springer.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kuhn, M., & Wickham, H. (2020). Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. *Boston, MA, USA.*

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, *116*(10), 4156–4165.

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. *arXiv preprint arXiv:1703.06856*.

Lambrecht, & Tucker. (2019). Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, *65*(7), 2966–2981.

Lee, & Riek. (2018). Reframing assistive robots to promote successful aging. *ACM Transactions on Human-Robot Interaction (THRI)*, *7*(1), 1–23.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc.

Mason, L. (2015). 'I disrespectfully agree'. The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, *59*(1), 128–145.

McClurg, S. D. (2006). The electoral relevance of political talk. Examining disagreement and expertise effects in social networks on political participation. *American Journal of Political Science*, *50*(3), 737–754.

McCright, A. M., & Dunlap, R. E. (2011). The politicization of climate change and polarization in the american public's views of global warming, 2001–2010. *The Sociological Quarterly*, *52*(2), 155–194.

Mitchell, T. (1980). *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research.

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, *21*(7), 959–977.

Monroe, K. R., Hankin, J., & Vechten, R. B. (2000). The psychological foundations of identity politics. *Annual Review of Political Science*, *3*, 419–447.

Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using twitter bots. *PloS one*, *12*(9), e0184148.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649.

Nagler, J., & Tucker, J. (2015). Drawing inferences and testing theories with big data. *Political Science and Politics*, *48*(1), 84–88.

Neunhoeffer, M., & Sternberg, S. (2019). How cross-validation can go wrong and what to do about it. *Political Analysis*, *27*(1), 101–106.

Norrander, B. (2018). The nature of crossover voters. In R. G. Boatright (Ed.), *Routledge handbook of primary elections* (pp. 119–132). Routledge.

of Governors of the Federal Reserve System (US), B. (2007). *Report to the congress on credit scoring and its effects on the availability and affordability of credit*. Board of Governors of the Federal Reserve System.

Okdie, B. M., & Rempala, D. M. (2019). Brief textual indicators of political orientation. *Journal of Language and Social Psychology*, *38*(1), 106–125.

Parekh, B. (2019). Liberal democracy and national minorities. *Ethnocentric political theory* (pp. 115–129). Springer.

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin UK.

Partheymüller, J., & Schmitt-Beck, R. (2012). A "social logic" of demobilization. The influence of political discussants on electoral participation at the 2009

German federal election. *Journal of Elections, Public Opinion and Parties*, *22*(4), 457–478.

Powers, E. (2017). My news feed is filtered? *Digital Journalism*, *5*(10), 1315–1335. https://doi.org/10.1080/21670811.2017.1286943

Qureshi, B., Kamiran, F., Karim, A., Ruggieri, S., & Pedreschi, D. (2020). Causal inference for social discrimination reasoning. *Journal of Intelligent Information Systems*, *54*(2), 425–437.

Rader, E., & Gray, R. (2015). Understanding user beliefs about algorithmic curation in the facebook news feed. *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 173–182.

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.

Reed, A., Whittaker, J., Votta, F., & Looney, S. (2019). Radical filter bubbles: Social media personalization algorithms and extremist content. *London: Global Research Network on Terrorism and Technology*.

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, *28*(1), 112–133.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., & Meira Jr, W. (2020). Auditing radicalization pathways on youtube. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 131–141.

Ribeiro, Singh, & Guestrin. (2016a). "why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 1135–1144.

Ribeiro, Singh, & Guestrin. (2016b). 'Why should I trust you?' Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Sanders, L. M. (1997). Against deliberation. *Political theory*, *25*(3), 347–376.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, *22*, 4349–4357.

Schmitt-Beck, R., & Partheymüller, J. (2016). A two-stage theory of discussant influence on vote choice in multiparty systems. *British Journal of Political Science*, *46*(2), 321–348.

Shikano, S. (2019). Hypothesis testing in the bayesian framework. *Swiss Political Science Review*, *25*(3), 288–299.

Sides, J., Tausanovitch, C., Vavreck, L., & Warshaw, C. (2020). On the representativeness of primary electorates. *British Journal of Political Science*, *50*(2), 677–685.

Singh, Singh, A., Khan, A., & Gupta, A. (2021). Developing a novel fair-loan-predictor through a multi-sensitive debiasing pipeline: Dualfair. *arXiv preprint arXiv:2110.08944*.

Slapin, J. B., & Kirkland, J. H. (2020). The sound of rebellion. Voting dissent and legislative speech in the uk house of commons. *Legislative Studies Quarterly*, *45*(2), 153–176.

Statista. (2019). Daily social media usage worldwide 2012-2019. http://tiny.cc/ecmbtz

Sterling, J., Jost, J. T., & Bonneau, R. (2020). Political psycholinguistics. A comprehensive analysis of the language habits of liberal and conservative social media users. *Journal of Personality and Social Psychology*, *118*(4), 805–834.

Stone, M. (1978). Cross-validation: A review. *Statistics: A Journal of Theoretical and Applied Statistics*, *9*(1), 127–139.

Stone, M. [Mervyn]. (1974). Cross-validation and multinomial prediction. *Biometrika*, *61*(3), 509–515.

Sunstein. (2017). *# republic*. Princeton university press.

Sunstein, C. R. (2017). *# republic: Divided democracy in the age of social media*. Princeton University Press.

Tajfel, H. (1981). *Human groups and social categories*. Cambridge University Press.

Tewksbury, D. (2003). What do americans really want to know? tracking the behavior of news readers on the internet. *Journal of communication*, *53*(4), 694–710.

Theodoridis, A. G. (2017). Me, myself, and (I), (D), or (R)? Partisanship and political cognition through the lens of implicit identity. *Journal of Politics*, *79*(4), 1253–1267.

Thornburg, M. P. (2014). Party registration and party self-identification. Exploring the role of electoral institutions in attitudes and behaviors. *Electoral Studies*, *36*, 137–148.

Tibshirani. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Trielli, D., & Diakopoulos, N. (2019). Search as news curator: The role of google in shaping attention to news information. *Proceedings of the 2019 CHI Conference on human factors in computing systems*, 1–15.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group. A self-categorization theory*. Blackwell.

Tutt, A. (2017). An fda for algorithms. *Admin. L. Rev.*, *69*, 83.

Twitter. (2017). Using Deep Learning at Scale in Twitter's Timelines. http://tiny.cc/mcmbtz

Twitter. (2019a). About your twitter timeline. *Twitter Help Center*. http://tiny.cc/fcmbtz

Twitter. (2019b). Improving engagement on digital ads with delayed feedback. http://tiny.cc/lcmbtz

Twitter. (2020a). Transparency around image cropping and changes to come. http://tiny.cc/kcmbtz

Twitter. (2020b). What Twitter learned from the Recsys 2020 Challenge. http://tiny.cc/jcmbtz

TwitterGov. (2016). Tuesday was the most-tweeted election day ever: 75+ million global election-related tweets sent through 3am et as trump claimed victory. http://tiny.cc/hcmbtz

van der Voort, H., Klievink, A., Arnaboldi, M., & Meijer, A. (2019). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? *Government Information Quarterly*, *36*(1), 27–38. https://doi.org/10.1016/j.giq.2018.10.011

Vapnik, V. (1991). Principles of risk minimization for learning theory. *NIPS*, *91*, 831–840.

Vapnik, V. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, *10*(5), 988–999.

Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021). Mitigating bias in machine learning for medicine. *Communications medicine*, *1*(1), 1–3.

Wasserman, L. (2015). *The role of assumptions in machine learning and statistics: Dont drink the koolaid* (tech. rep.). Technical report, Carnegie Mellon University, 2015. 8.

Webb, G. I., & Zheng, Z. (2004). Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. *IEEE Transactions on Knowledge and Data Engineering*, *16*(8), 980–991.

Weber, M. S., & Kosterich, A. (2018). Coding the news: The role of computer code in filtering and distributing news. *Digital Journalism*, *6*(3), 310–329.

Weizenbaum, J. (1976). Computer power and human reason: From judgment to calculation.

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, *1168*(2), 022022.

Young, I. M. (2002). *Inclusion and democracy*. Oxford University press on demand.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment disparate impact. *Proceedings of the 26th International Conference on World Wide Web*.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. *CoRR*, *abs/2106.11342*.

Zhou, Kantarcioglu, M., & Clifton, C. (2021). Improving fairness of ai systems with lossless de-biasing. *arXiv preprint arXiv:2105.04534*.

Zhou, Sap, M., Swayamdipta, S., Smith, N. A., & Choi, Y. (2021). Challenges in automated debiasing for toxic language detection. *arXiv preprint arXiv:2102.00086*.

Zou, J., & Schiebinger, L. (2018). Ai can be sexist and racist—it's time to make it fair.

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of information technology*, *30*(1), 75–89.

Zuiderveen Borgesius, F., Trilling, D., Möller, J., Bodó, B., de Vreese, C., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*, *5*(1).