Public opinion without polls:

Investigating the feasibility of Twitter-based election forecasts

NIKLAS M. LOYNES

A thesis for the degree of Doctor of Philosophy
submitted to:
Faculty of Humanities
School of Social Sciences,
The University of Manchester
August 16, 2021

Contents

List of Figures

List of Tables

Abstract

This thesis investigates whether, and if so, how digital trace data taken from the social media platform Twitter can be used as a valid and reliable basis for measuring public opinion. While Twitter data has been used in a range of spheres to measure public opinion, in the political research sphere work has typically centred on forecasting the outcome of elections. This functions as a predictive exercise in its own right, but can also be framed as providing a test for the robustness of measures derived from Twitter data as a new proxy for voter opinion: election outcomes offer a 'ground truth' against which bias in these data can be assessed and the extent of error measured. Results to date have been mixed, with studies showing varying levels of accuracy in matching their estimates of vote intention against polls and/or electoral outcomes. So far, however, none have consistently achieved the standards of accuracy, reproducibility and reliability that would make them an unquestioned alternative to surveys, thereby offering ample scope for new research, which this thesis contributes to.

The research documented in this thesis is guided by the assumption that the sheer amount of political content on Twitter provides enough pertinent and varied information to measure public opinion, but the tools necessary for achieving this reliably and at scale do not yet exist. The core goals of this thesis are exploring which methods are required, and contributing to their development. This is approached in three papers. First, 'Finding Friends' demonstrates a new approach to estimating any given Twitter user's home location. Second, 'Understanding Political Sentiment' benchmarks existing approaches to measuring public opinion by extracting sentiment from tweets, as well as a new approach to estimating public opinion metrics derived from hand-labelled and machine-propagated sentiment scores, while applying twelve aggregate vote share prediction models in a comparative framework in three US 2016 presidential primaries. This paper, as well as third paper employ samples of Twitter users generated using the geo-locating algorithm outlined in paper 1. Paper 3, 'Listening in on the noise' introduces a novel approach to estimating individual-level political preferences using distant supervision and Machine Learning. This is applied on two distinct, geo-located samples of users, one selected based on indication of previous voting, one selected randomly, in order to trace public opinion in the 2018 US midterm elections, both nationally and in the four most populous states.

Each of my papers marks a significant contribution to its particular issue-area in the field of Twitter-based public opinion research. Paper 1 adds a reproducible geo-locating pipeline with an open source software package. Paper 2 adds evidence to the evaluation of the usefulness of sentiment analysis on tweets for public opinion measurement. Paper 3 introduces an approach to estimating individual-level political preferences, as well as highlighting the impact sampling decisions have on research outcomes. The best models for predicting vote shares in papers 2 and 3 achieve mean errors on par with opinion polling, and offer, through their application in diverse election scenarios and by following a theory-grounded and reproducible framework, a strong contribution to existing practice in the field. Furthermore, the ability to estimate individual users' home locations significantly improves sampling capabilities for researchers employing Twitter data across fields.

Keywords: Twitter, elections, public opinion, forecasting, American politics, computational social science

Supporting Materials

This thesis, being in the field of computational social science, relies heavily on specially designed software to gather data, process and arrange data, analyse data, generate graphs and figures and to automate recurring tasks. For the purpose of transparency and reproducibility, I share all required code which was written in order to complete these processes. Besides the code, this thesis also relies on data collected from the Twitter API. In compliance with Twitter's terms of service, I do not share this data publicly, but only provide it upon request to collaborators and examiners. Some of the links below may require me to provide access to repositories, if the publication process for individual papers does not yet allow for them to be public.

- Paper 1, 'Finding Friends':
    1. Complete software package allowing for the geo-location of Twitter users, github.com/nikloynes/geolocation
    2. Replication code for the analysis presented in the paper github.com/nikloynes/geo_locating-paper_code
- Paper 2, 'Understanding Political Sentiment'
    1. Replication code for the analysis presented in the paper: github.com/nikloynes/sentiment_paper_rework
- Paper 3, 'Listening in on the noise'
    1. Replication code for the analysis presented in the paper: github.com/nikloynes/midterms_paper

I can confirm that, at the time of submission, all code shared here compiles as intended and produces the same results as presented in the papers. However, given the dynamic nature of internet-enabled computing, and the fact that all of the code presented here imports third-party, open-source libraries which are subject to changes, I cannot guarantee that the code will run without modifications, or on any system that is not the system I ran this code on. Therefore this software is presented without any warranties of any kind, and should be understood as complementary to the thesis. Furthermore, this software is hosted on www.github.com. I cannot guarantee the permanence of links, or the continuing existence of this company, therefore the enduring functionality of links cannot be guaranteed by this author.

Declaration

I, Niklas Loynes, hereby certify that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

There are a range of people who deserve explicit thanks for helping me along the way of the journey to submitting a PhD, I would like to mention them here.

First and foremost, the biggest thanks goes to the best supervisory team one could possibly imagine: Rachel Gibson, Marta Cantijoch Cunill and Mark Elliot, who really made this thesis possible with their advice, insights and encouragement. Each of them contributed to this thesis in their own way, and I learned a huge deal from all three. Specifically, I have to thank Rachel Gibson, who was always ahead of the curve, and managed my entire PhD in the kindest, most thoughtful of ways. I am certain that I would not have been able to complete this thesis had it not been for Rachel's perseverance and patience.

Christa Rehberger-Loynes, without whose support and encouragement I would never have made it anywhere close to considering a PhD.

Amsha Kalra, who helped me with everything - be it motivation, support and encouragement, be it prettifying my ugly, grey graphs, be it cooking delicious meals while I debugged my code, or proofreading my papers and proposals.

Judy and Iggy Barton, who financially supported me when I most needed it.

Sarah Alexander and Yosh Wakeham, for proofreading my proposals at the 11th hour, and always offering constructive feedback, ideas, encouragement and coding support.

Benedikt Neutard and Niklas Schöfer, who were always up for a match of Fortnite.

Olatz Ribera, for housing me in Barcelona when I was there for a conference.

Paul James and Matt Thompson, for helping me write an abstract for APSA (which got accepted!) in the most intoxicated of states.

Paul James, for always offering me a home away from home in Manchester, for making my move to New York possible with his help and for proofreading my work.

Ben Sessions, for always offering me a home away from home in London, where I wrote a considerable amount of this thesis, and for proofreading my work.

Rachel Alexander, for giving me a desk at LSE, where I wrote much of the third paper in this thesis.

Julia Maclachlan, without whom I would probably never have learned what the letters 'P', 'h' and 'D' stand for, let alone have them be something I seek to achieve.

Jon Las Heras, Ben Sessions, Sara Van Goozen and Wen-Chin Lung, for inviting me to share their home when I really needed it, and showing me what it takes to be a PhD student.

Jon Las Heras and Ilias Alami, thanks to whom I realised that there is more to science, research and critical inquiry than rigid empiricism.

Billy Christmas, who made many of my Manchester nights memorable... or not.

Jonathan Nagler, for being my unofficial fourth supervisor, imparting a gigantic bowl of wisdom upon me, pushing me to become a better coder and researcher, and helping me out

when I needed it most.

The kind people who contribute to places such as GitHub and StackOverflow, without whom I would not even know how to write ’Hello World’.

John Bryden, for baring with me while I juggled multiple projects.

Pablo Barbera, for sharing vital and useful data with me.

A special thanks goes out to all the people at SMaPP, specifically:

Andreu Casas, for being a friend and mentor throughout my time there.

Greg Eady, for always offering help and advice, and always being available to fix my code, and for a great week in Barcelona.

Cody Buntain and Frido Linder, for helping me figure out how to use MySQL efficiently.

Yvan Scher, Leon Yin and Nicole Baram, for helping me become a better coder.

Haohan Chen, for helping me figure out how to do Python machine learning correctly.

Zeve Sanderson, for always sorting anything out that needed sorting out, and being the most reliable manager one could possibly hope for.

MaryJane Dumankaya, who went above and beyond to help me out with visa stuff.

Tasha Gordon, Megan Brown and Andy Guess, for tolerating me as an office-mate.

Further thanks goes to everyone at PSSRU Manchester from 2014-16, specifically:

David Challis and Angela Worden for believing in my skills as a researcher, and encouraging me to pursue a PhD, and further letting me prioritise my studies when I needed to. Paul Clarkson, for proofreading my proposals, and giving me advice and encouragement.

Thanks also to the folks who work in and around the School of Social Sciences at the University of Manchester, specifically

Bernadette O’Connor, Ann Cronley and Kimberly Hulme, for always figuring stuff out for me, and not letting bureaucracy stand in the way.

Liam Shields, without whose help in getting funding, I would never have been able to start my PhD.

Olga Onuch and Maria Sobolewska, for being excellent programme directors.

Sophie Woodward, for being such a kind and good-natured, encouraging completion support person.

Ed Fieldhouse and Maria Sobolewska, for providing incredibly useful feedback as my annual reviewers.

Paul Johnston and the whole team at Research IT, for providing me with an impressive amount of compute.

Magnus Rattray and the UoM Data Science institute, for providing the funding necessary for me to go to New York.

Kaicheng Yang, David Carlisle and Ulrike Fischer for saving my everything at the 11th hour, 59th minute and 59th second by identifying why my LaTeX code wouldn’t compile.

And last, but certainly not least, Rob Ford and Luke Sloan, for examining this work and giving useful and welcome feedback.

For Harry, who would have enjoyed reading this.

Chapter 1

Introduction

In this thesis, I investigate the feasibility of measuring public opinion using digital trace data generated as a consequence of expressed human behaviour on the social media platform Twitter. Given the ability to precisely compute measurement error, this pursuit is applied in the form of predicting election results using data gathered from Twitter. In contrast to existing research in this vein, the election forecasting element of this research is a means to furthering the understanding of how public opinion can be measured, rather than an end in itself. This is the case as elections offer a verifiable outcome quantity, based on which the accuracy of a given model can be assessed. This research can be broadly situated at the intersection of political methodology, social data science/computational social science, and public opinion research.

My contributions to the field are primarily methodological: I develop, outline and empirically test several novel approaches to extracting public opinion-relevant signals from Twitter data, while addressing shortcomings of previous research in the area. In doing so, I add to the theoretical and conceptual understanding of how Twitter data can be used in public opinion research, pursuing an agenda of making tweets ′fit for purpose′ for a mature discipline of Twitter-based public opinion research. My contributions are presented in three papers, each of which examine a different, yet equally important issue area within the field. First, I present a scaleable method for geo-locating Twitter users, which allows for more reliable sampling, weighting and de-biasing, in ′Finding friends′. Second, I provide an in-depth examination of the role of computational sentiment analysis in Twitter-based public opinion research, empirically benchmark existing approaches, and present novel techniques for extracting and further utilising sentiment-labelled tweets in ′Understanding Political Sentiment′. Third, I present a highly accurate method for estimating individual-level political preferences from tweets using machine learning and distant supervision, in ′Listening in on the Noise′. Notwithstanding the papers′ nature as stand-alone pieces of science, they build upon one another, and leverage methods developed therein - most prominently the geo-locating algorithm - to enhance analysis, and thereby produce a comprehensive, integrated body of work. These contributions add most pertinently to an emerging field of research in political science and computational social science which is broadly concerned with measuring public opinion through Twitter data, with papers focusing on forecasting election results from tweets (Tumasjan et al., 2010; Jungherr et al., 2012), forecasting election results using tweets classified for their sentiment (Ceron et al., 2014) or with methods aimed at measuring latent, political and public opinion-salient individual-level traits such as ideology (Barberá, 2015).

In order to situate this research, it is useful to first describe the environment from which it draws: the social media landscape of the maturing 21st century. Social media has developed from a range of fringe, early-adopter communication tools into a key structuring force of global digital society in 2021. Humans' digital inter-personal communications now take place on WhatsApp, iMessage or Facebook Messenger; individuals record and stream their lives in real time on Instagram, Snapchat and Twitch; scheduled TV programming is fast being overtaken by on-demand, social video content platforms such as YouTube, while news is disseminated, discussed and *made* on Twitter.

Beyond the re-structuring of the global media ecosystem, the shift from humans as passive media consumers to producer-consumers of dynamic, interactive, multi-media and multi-platform content brings with it a giant stream of individual-level digital trace data - be it content produced by individuals, such as posts, comments, articles, videos, images and live broadcasts, or content- and user-level metadata. These digital trace data are primarily used by the platform companies on which they were generated to advance their business interests, e.g. by targeting users with personalised advertising or content generated by other users expected to maximise engagement, and thus platform revenue.

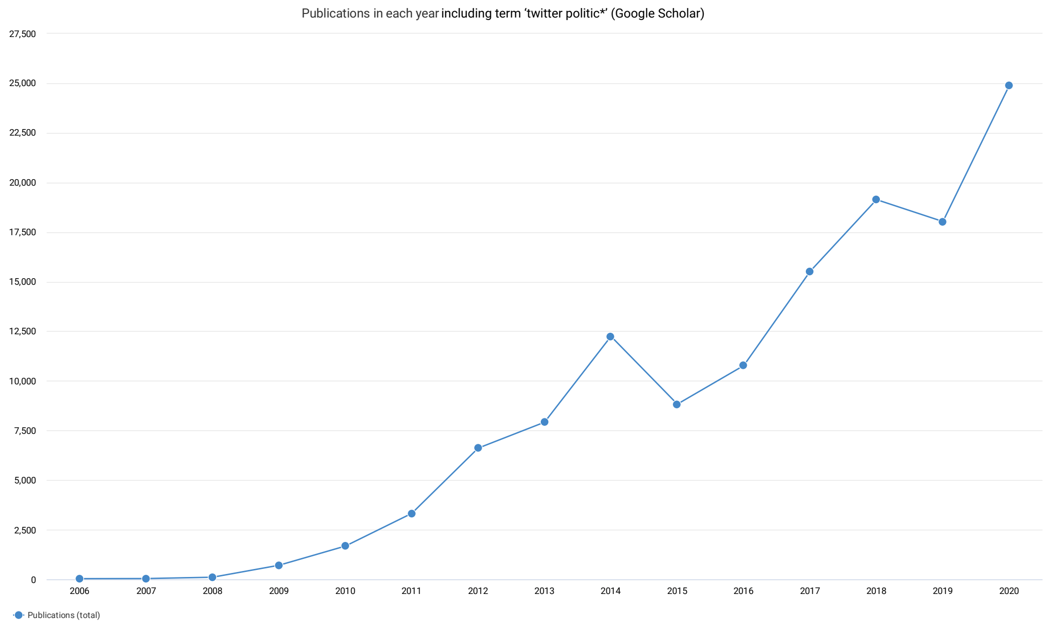
Access to user-level digital trace data is governed by social media platform companies, and is becoming increasingly more restricted for the vast majority of platforms: Facebook limits access to posts and comments on public groups, while Instagram and WhatsApp (both of which are also owned by Facebook) provide no access to data beyond that of the authenticating user. In practice, this means that most social science research availing itself of social media digital trace data relies on Twitter, which, through its *Search* and *Streaming* Application Programming Interfaces (APIs[1]) allows easy access to large-scale tweet and user-level data, which can be collected dynamically, i.e. as they are published (Streaming API) or retrospectively (Search API).

Individual-level digital trace data from social media platforms have also had an enormous impact on the social sciences, and crucially to this thesis, the way in which social science is conducted: digital trace data provide a cost-effective approach to investigating research questions relevant to human behaviour at scale, be it through analysing digital trace data for answering research questions concerning offline or online phenomena, recruiting research participants on social media, or disseminating research findings with previously hard-to-reach audiences and building global online research communities. This is illustrated in Figure 1.1, which show the growth of academic publications containing relevant keywords, in this case *'twitter politic*'*, since 2010[2].

Equating Twitter and its user-base with 'the online population', or, more than that, target *offline* populations would be hugely erroneous: not only is Twitter differently popular in different locales around the world, it is also disproportionately used by young, well-educated, left-leaning and urban users (Wojcik and Hughes, 2019). In other words, Twitter's user-base is likely to be significantly less representative of most offline populations than e.g. Facebook's, which has significantly more users and is more of a catch-all platform than Twitter, with more users from different demographic backgrounds and a myriad of different products attracting a variety of users, as opposed to Twitter's singular product. While the incongruent nature of

[1] An API is provided by a service/application/website etc. to enable developers to access data and other resources, which the developer can in turn incorporate into their own applications

[2] This graph was generated using app.dimensions.ai on 14/02/2021.



Source: https://app.dimensions.ai
Exported: February 14, 2021
Criteria: Text - 'twitter politic*' in full data.

Figure 1.1: Number of academic publications containing ‘twitter politic*’, retrieved from Google Scholar (using app.dimensions.ai)

Twitter's population versus target offline populations is relatively unproblematic in research scenarios focused on Twitter users, it is important to note that Twitter is reportedly home to an exceedingly large amount of automated, non-human accounts, commonly referred to as 'bots'. While there is no authoritative figure on the prevalence of bots, research suggests that "between 9% and 15% of active Twitter accounts are bots" (Varol et al., 2017, p. 280), and "[a]n estimated two-thirds of tweeted links to popular websites are posted by automated accounts" (Wojcik et al., 2018). Reliably discriminating between humans and bots at scale is a complex task which further complicates Twitter-based social science research. Finally, the minimal nature of Twitter's user profile - the only required field is a free-form text 'screen name' - makes the identification of relevant user-level demographic attributes, which could be used to weight and de-bias samples of tweets or users, or reliably classify accounts as bots or humans, very complex.

Nonetheless, Twitter is home to around 330 million monthly active users (Statista.com, 2019), who together create a staggering 500 million tweets every single day (InternetLiveStats.com, 2019). A tweet can range in informational depth from a few, potentially hard-to-interpret symbols to 240-character prose complete with images, videos, and URLs; at times strung together into 'threads' which rival traditional journalistic content, albeit without any editorial restrictions. Furthermore, evidence indicates (Goel et al., 2010; Colleoni et al., 2014) that 10% of tweets are of a political nature. In short, this means that the emergence of Twitter has provided political scientists with the *largest ever* resource of individual-level data pertaining to peoples' views on political issues, which, unlike survey data, are generated through voluntary behaviour on the side of their creators. Hence it follows - notwithstanding the inherent issues of representativeness, (demographic) bias or bots associated with the Twitter population - that this data stream offers a unique opportunity for innovation in the field of public opinion research. On the pathway towards a Twitter-based public opinion infrastructure, there are however a large number of problems to solve in order for the endeavour to be a fruitful one. This thesis contributes to the field by both illuminating and better understanding such obstacles, and further, developing methodological approaches for overcoming them.

Surveys: declining response rates and increasing costs

At the centre of this thesis stands the puzzle of how observable records of politically salient digitally-mediated human behaviour - tweets - relate to a notion one may refer to as "public opinion". The underlying assumption being that if public opinion is an aggregation of the political views, attitudes, values and preferences of a defined group of individuals, then Twitter, the largest collection of individuals' documented political views, attitudes, values and preferences in human history, should provide an invaluable resource for the mapping of public opinion in the 21st century. I argue that it is of considerable importance for public opinion, in the fine-grained diversity one should expect of it in the modern world, to be reliably and transparently parseable, as functioning democratic governance draws a large amount of its legitimacy from the belief that it is responsive to the wishes and aspirations of those individuals it governs.

One may call into question the necessity of this pursuit, given the existence of an established industry and scientific tradition devoted to measuring public opinion, namely polling and survey research. However, there are several recent developments in this field which highlight the necessity of investigating alternatives. First, survey response rates have been dramatically declining in the recent past (Groves, 2011; Kennedy and Hartig, 2019), meaning that obtaining a reliably representative sample necessary for accurate measurements of public opinion is becoming increasingly difficult and costly. Second, polling, a regular application of scientific surveys to measure public opinion, typically on political and electoral issues, has recently produced several well-publicised failures (assuming the goal of pre-election polls is to closely approximate the eventual election result), for example the incorrectly forecast outcomes of the UK's 2016 EU-membership referendum, the 2015 UK general election, the 2016 US Presidential election or the 2017 UK general election. While the reasons for these failures may partially be down to the decline in survey responses, as well as a general shift away from landline telephones (they allow for easy and reliable geographic sorting of respondents), it also appears that the commercial nature of much of applied survey research, which generates most of the survey-derived findings disseminated by the media, but also those further analysed by academic survey scientists, plays a part in this. Its proprietary approaches to recruitment, sampling, interviewing or weighting make the identification of systemic issues by independent actors, in the style of e.g. the academic peer review system or the open source model for software development, difficult without willing cooperation from the side of the commercial interests. In other words - a pollster is thoroughly within their rights to disclose nothing about their design decisions to the public, and it is up to either media outlets or savvy individuals to decide which specific polls and pollsters to trust. However, even assuming sophisticated modern approaches to sample stratification, imputation and synthetic data generation in lieu of lacking observations *do* produce adequate snapshots of public opinion, it is possible that statistical approaches and modelling can only go so far if the key quantity of interest - regular people sharing their opinions on pertinent issues - is becoming increasingly difficult and costly to obtain. This further ties into a third issue with the survey research paradigm: if public opinion is information essential for the health of a democracy, then it may not be desirable that the mechanisms used for identifying such information are essentially black boxes beyond public oversight. While this may not have been pertinent when multiple companies were providing equally reliable, comparably easily accessible information, in an ever-more consolidated industry where only few players have the ability to reliably measure public opinion (see e.g. Silver, 2014), I suggest that there is increasing need for transparency. It is however important to note that the survey paradigm is not *only* applied by for-profit, closed source actors - indeed, the most successful and reliable actors in the survey landscape are typically government agencies, or other publicly funded institutions, such as the European Social Survey, the British Social Attitudes Survey, Eurobarometer, national election studies, or, indeed, national census bureaus. However, I would argue that this highlights the immense cost and complexity associated with conducting ’good’ survey science.

The need for alternative approaches to quantifying public opinion goes beyond the critique of the closed source model of the polling industry, its recent bad track record, or the exclusivity and rarity of public-sector survey research. Humanity is currently finding itself in the midst of the ’information revolution’, where humans, the *users* of digital products and services, are producing giant amounts of data, documenting and cross-referencing most any behaviour, be it the swiping of a MetroCard on their morning commute, the purchase of a coffee before boarding the train, the subway WiFi they log on to, as well as all websites they visit while on this WiFi, potentially annotated with GPS coordinates and timestamps. These giant amounts of individual-level data are fuelling much of the growth (but not necessarily profitability) in the new digital economy, as they are used by the companies for whom they were generated to learn more and more about their users, and to - in some way or other - monetise this knowledge, be it by showing targeted advertising to said users or by selling user data to third parties. The way in which these data are monetised often relies on a large-scale, applied form of social science - with far less weight given to the importance of solid theory, but rather with companies’ bottom lines subverting core principles of rigorous, ethical scientific inquiry. Nonetheless - the availability of these vast amounts of data highlights that social scientific and statistical approaches are useful for predicting and mapping human behaviour, given the right data and the right computational capacities. So, given the immense cost of conducting surveys and the opportunities offered by observational, non-intrusive digital trace data created as by-products of routine human behaviour over elicited responses to questions many individuals may have never have had opinions on prior to being asked, I argue that it is likely that the study of public opinion will increasingly shift beyond the pure survey paradigm, and instead avail itself of the vast troves of digital trace data documenting human (political) behaviour to supplement survey research. However, I also argue that it is of urgent importance that the tools and the governance of the charting of public opinion through digital trace data must not be left to the for-profit companies who currently govern them. The degree to which the unintended consequences of business interests of companies dealing in digital trace data (e.g. Facebook, Twitter, Google) have already enabled an environment of historically low trust in formerly powerful institutions of government and media, and seemingly puts much of the established social order into jeopardy, indicates that these actors would likely use their power and command of digital trace data not for the public good, but rather for their own. Hence, it is of utmost relevance that academics and social scientists further the development of a modern, transparent and open, digital trace data-led framework for the study of human behaviour and public opinion.

The Case for Twitter

At the time of writing, the default data source for social science research using digital trace data is Twitter. While it is currently the only large-scale social media platform offering generous data access to developers and researchers and thus is the default by necessity, it is also uniquely situated for playing this role. First, it is public on the web, and users share content on the platform in full understanding that this is the case. This is in stark contrast to the other dominant social media platforms in 2021, such as Facebook, Instagram or Snapchat, all of which operate a product which gives its users the semi-implied, semi-stated, understanding that they are operating in quasi-private networks only including other individuals they have actively connected with[3]. Hence, it would be unethical to collect individual-level user data from said platforms even if it were available to researchers, as content is likely to have been shared in the understanding that it is viewable only by a readily comprehensible audience. Twitter content on the other hand is typically shared with a quasi-public audience in mind[4]. However, Twitter's use and reach goes beyond merely being a platform where individuals can share content. Rather, it is important to emphasise the *Media* in social media when discussing the site, and explaining why it is connected to public opinion. Namely, Twitter as a platform provides endless media (and thus political) content for individuals to consume. This transcends the number of active, or indeed registered users on the platform, as modern-day political communication is increasingly taking place on Twitter. Take, for instance, the 45th President of the United States: even if any given individual may have wanted to avoid Donald Trump's tweets, many of them invariably filter through Twitter and dictate the news cycle on all media platforms, social or otherwise. Hence, the reach of news and current events content which is generated primarily through Twitter largely exceeds the number of registered or active users, and it therefore follows that content on the platform has a wide-reaching effect on public opinion. Besides the reach of Twitter-based content beyond the platform, it also serves as a space for individuals as media consumers and *relayers*. The site's retweet feature allows individuals to share content with their followers (and, again, the world at large), a core element of the phenomenon of virality (see e.g. Hansen et al., 2011). Furthermore, the platform allows for the emergence and cultivation of opinion leaders (see e.g. Katz and Lazarsfeld, 1955)- individuals whose views are trusted and adopted by their followers. Additionally, the site is not only useful for sharing of content and opinion, but also for (political) communication. Indeed, it has never been as easy for individuals to engage in conversation with others, sometimes the aforementioned opinion leaders, sometimes just other random individuals. While this sometimes produces engaging debate and allows for the dissemination and adoption of new viewpoints, it has also led to a large-scale increase in adversarial interactions - oftentimes due to disputes of a

[3] This is subject to the authenticating user's profile configuration. Instagram accounts, for instance, can be set to be private or public, whereas Facebook accounts have more complex, multi-layered network scope settings. The effectiveness of such user settings is further complicated by the ongoing merging and interplay among Facebook's differently branded products

[4] This definition excludes the small minority of Twitter users who set their account to "private", thereby only displaying their content to users they have granted access to it - and thus excluding researchers as well

political nature. Regardless of the nature of such inter-user interactions, they offer a further facet of relevant political communication to the range of data to be studied. And, finally, and arguably most importantly, Twitter offers an arena for anyone to simply share their opinions on anything, regardless of whether they result in engagement from others or not. Crucially, for researchers at least, these opinions are now documented, and can be studied.

As stated above, however, Twitter data is far from ideal for the purpose of conducting public opinion research. Hence, it is important to first identify some of the immediate issues facing those wishing to mine tweets for public opinion (see e.g. Mislove et al., 2011; Gayo-Avello, 2013; Klašnja et al., 2017):

1. *The Twitter population is not congruent with most any target offline Population*. In practice, this means that Twitter offers a non-random sample of the general (global) population, and any inferences made to populations outside the Twitter population from Twitter data will not be accurate.

2. The exact composition of the Twitter population - demographic, geographic, ethnic, religious, country of origin, etc. - is unknown. Twitter data does not provide any information regarding such individual-level attributes.

3. There is no guarantee that such individual-level attributes can be accurately inferred, computationally, interpretatively or otherwise.

4. Indeed, there is no guarantee that a Twitter user can be definitively classified as human.

5. Even if such characteristics could be reliably inferred, they do not necessarily offer a heuristic for ensuring a given Twitter user is included in a desired sampling frame of Twitter users, e.g. Democrats from California.

6. Tweet collection using the Streaming API requires *a priori* definition of relevant keywords (unless researchers want to collect a sample of at most 1% of all tweets[5]). There is no guarantee that such a researcher-defined keyword list would be comprehensive and exhaustive, and thereby has the potential of biasing research findings.

7. Out of the entirety of the Twitter population, or any issue or group-specific sub-population ('political Twitter', 'Baseball Twitter', 'Underground Hip-Hop Twitter'), the number of users who dominate conversation are a small proportion of that sub-population. How can we be sure that we are actually studying who and what we wish to study if conversation is dominated by its 'loudest' participants?

8. The motivation for tweeting, or any other user-level interaction on Twitter, is unknown - users may accidentally be hitting the *like* or *retweet* buttons. In the case of ascribing public opinion-relevant signals to individual-level online behaviour, this has the potential of resulting in falsely ascribed intentions and signals.

9. For most research areas of interest, simply counting the occurrence of *n* (e.g. tweets containing a keyword, retweeting a post, following an account) is not sufficient for establishing, recording and categorising a behaviour of interest. Rather, this requires a degree of interpretation.

[5] There is ample academic debate regarding whether the 'public sample stream' is a random sample of tweets (Morstatter et al., 2013), given that Twitter does not publish the exact parameters by which it is generated

10. Often, we do not know how to interpret (political) language on Twitter, even if this process involves careful reading, labelling and coding. Hence, any interpretation of text data at scale is accompanied by uncertainty and bias.

While this list is by no means comprehensive, it touches on some of the core issues facing social scientists working with Twitter data. The good news is that there are ways of addressing these issues. Furthermore, I argue that further development in illuminating these problem areas of working with Twitter data can bring with it the emergence and development of a robust theoretical framework for understanding public opinion, political behaviour and politics through Twitter, guided by stronger empirical knowledge of the environment that is being studied. This thesis aims to contribute to this quest, with a strong focus on the inference of individual-level demographic user attributes and the systematic and reliable extraction of politically salient information pertaining to individuals' political views and attitudes from Twitter-derived text data.

Outlining a research agenda

> *"Much methodological and theoretical work has yet to be done to integrate these data sources fully into the social sciences. Most of the available research focuses on the presentation of single-shot case studies, which claim to show the relationship between some metrics of attention online with some metrics of political behavior. Any meaningful theoretical discussion of potential mechanisms between specific patterns in online trace data and political phenomena is missing from the literature. Thus, it is very difficult to assess whether or not case studies offer more than accounts of spurious correlations between otherwise unconnected variables. This unsatisfactory state of the field is due to the fact that most of the research analyzing political behavior through online trace data has been presented by computer scientists who work in a publication culture that favors short peer-reviewed conference papers."* (Jungherr, 2015, p. 5)

While I broadly concur with this assessment, I find it useful to add a distinction here, which, perhaps, has evolved in the time since its release: there are different kinds of goals that researchers follow when conducting research with digital trace data. On the one hand stand the examples outlined by Jungherr: Twitter data as a quasi-mirror of the real world, not a biased and incomplete representation of it. Then, *ad-hoc* and inductive methods are applied in order to 'prove' something, but underpinnings of why this might be are under-theorised. This is the arguably not-so-rigorous part of what I call *research using Twitter to study politics*. However, there is another approach, which studies political behaviour *on Twitter*. This might involve bots, disinformation, fake news, etc. In this case, with clearly empirical questions and much lesser of a need for operationalisations and proxies, it seems unfair to suggest that this research is under-theorized and not sufficiently rigorous. This thesis is strongly within the first column - studying politics with Twitter. I am interested in how Twitter reflects politics in the real world (substantively), and how the large amounts of data that we can collect from it can add be used to study it. So, the fundamental objective of this thesis is to improve the rigour of Twitter-based social science research by introducing, testing and benchmarking novel approaches that address some of the shortcomings with the discipline, as they exist at the time of writing (and itemised above). This way, the goal is to add knowledge to the field which can improve the reliability and validity of inferences on politically salient human behaviour and attitudes generated from Twitter data. By doing this, I further seek to provide a stronger foundation and thorough justification for the overall aim of measuring offline public opinion with Twitter data.

For this, the purpose of examining the feasibility of reliable and reproducible measurement of public opinion from social media data and to contribute to the methodological advancement of the field, it is useful to define a specific scenario through which this can be investigated. Hence, I choose to focus on the forecasting of election results in developed democracies. Election forecasting is the process of using available and related data to estimate the eventual outcome of a multi-party/multi-candidate democratic election. While this is a valuable pursuit in its own right, election forecasting offers the unique opportunity of providing an unambiguously true outcome quantity that forecasting aims to measure as accurately as possible. In the social sciences, and in political science in particular, such true outcomes are rare - most of the time, research findings and evidence are considered true if they are supported by multiple sources or seen as intuitively true. However, having the ability of knowing the exact quantity that a model *should* predict allows for a thorough evaluation of the performance of the model, and further to use an established relationship to investigate questions beyond the core quantity of interest, given that relationships between dependent and independent variables can be confidently deduced. This has been the key contribution of the existing election forecasting literature in political science, which, starting in the 1970s, developed varied statistical models for forecasting election results with significant lead time prior to an eventual election date using a range of variables, such as candidate approval ratings (see e.g. Lewis-Beck and Rice, 1982), or the state of the economy (see e.g. Fair, 1978; Lewis-Beck, 1990; Lewis-Beck and Stegmaier, 2007). Crucially, this research illuminated the importance of incumbency and the state of the economy on vote choice - factors which, at the time, were hotly disputed but are now considered quasi-canonical. In other words, these papers showed that while it is clearly interesting and useful in its own right to be able to reliably forecast election results without the need of regularly consulting polling data or even knowing the candidates' names, the case of scientific, statistical election forecasting added considerable evidence to our understanding of the key determinants of election outcomes, which in turn have greatly benefited other sub-fields.

In this thesis, I aim to use the scenario of election forecasting as a vehicle for furthering the understanding of how digitally mediated political behaviour relates to (offline) public opinion. If social digital trace data are the new go-to data source for political scientists, then it is of utmost importance to increase the understanding of what such data signify in a political context, and hence what can be deduced from them, but also which applications certain forms of digital trace data may be useful for, and which ones not. Furthermore, this pursuit can help illuminate our understanding of human (political) behaviour as mediated by social media platforms. When, why and how do individuals tweet (post) about politics? How do the structures of human relationships as embodied on social media platforms shape the spread of information, and the shaping of individual- and group-level beliefs, opinions, preferences and positions, be it in a political or a general context? More specifically, applied to the political-electoral context: how does online support of individuals, organisations/institutions or causes - embodied through friend-following relationships or through user-to-user interactions - measurably translate to the offline world, if at all? While any single one of these questions is clearly too broad and complex to be investigated in depth in this thesis, they are, for the most part, both mostly unanswered and subject to change over time.

Even a minor, incidental contribution to the understanding of the underlying patterns that explain the way these new technologies shape human behaviour and society more broadly can be of enormous value to the social science community, regulators and policy-makers, and individuals. So, in keeping with the example set by the existing election forecasting literature and its achievement of illuminating structural determinants of democratic elections, this research also strives to illuminate the broader conceptual meaning of digital trace data in social science research and digital society more broadly. This goes beyond the understanding of tweets as data, but rather of tweets as artifacts of human communication, and thus, to a certain degree, core components of human (political) behaviour in the digital age. Through this process, I seek to contribute to the toolkit of methods available to computational social/political scientists, but also to increase conceptual understanding of the available data and its underlying meanings.

The question that arises, then, is how to go about investigating such research questions? Moreover, how can *political* meaning be derived from large collections of 280-character tweets? Intuitively, Twitter data, especially those conveniently collected using the streaming API and filtered by targeted, pre-defined keywords, provide a simple metric by which the Twitter-wide popularity of (political) concepts, terms, parties, candidates, etc. can be measured - *keyword mention frequency*, both relative to other keywords, or a larger population of tweets forming a pertinent collection. Assuming that relevant keywords by which the Twitter stream is to be filtered are correctly and adequately defined (and even knowable) *a priori*, it is only a case of comparing the *volume* of given keywords or keyword groups against one another for the same time frame. This concept was applied to the problem of election forecasting in Tumasjan et al. (2010), a seminal paper for Twitter-based election research. The authors claimed to have forecast the German Bundestag election of 2009 using the volume of pre-specified keywords. However, Jungherr et al. (2012) showed that slightly modified keyword specifications - on the same data - produced greatly differing election forecasts, to the point where the widely insignificant German Pirate Party would have emerged victorious in 2009, were we to trust keyword volume as a predictor of election outcomes.

In order to illustrate the issue with simply measuring the volume of keywords in tweets, consider Figure 1.2. It depicts, one, the mention volume percentage of the keywords "democrat*" and "republican*" of Twitter's 1% sample stream[6] for 3 months leading up to each country-wide election in the USA between 2012 and 2018, as well as the relative difference in the observed occurrence of either keyword.

[6] Twitter provides a sample of all tweets that are published to the platform in real-time through its streaming API. This 'sample stream', sometimes referred to as the 'Spritzer', is set up to be a 1% sample of the entire stream, and Twitter claims that this sample is a random subset of all tweets. However, research indicates (see e.g. Morstatter et al., 2013; Steinert-Threlkeld, 2017; Paik and Lin, 2015) that this sample is not truly random, so broader conclusions regarding the nature of the platform and behaviour thereon should be approached with caution.

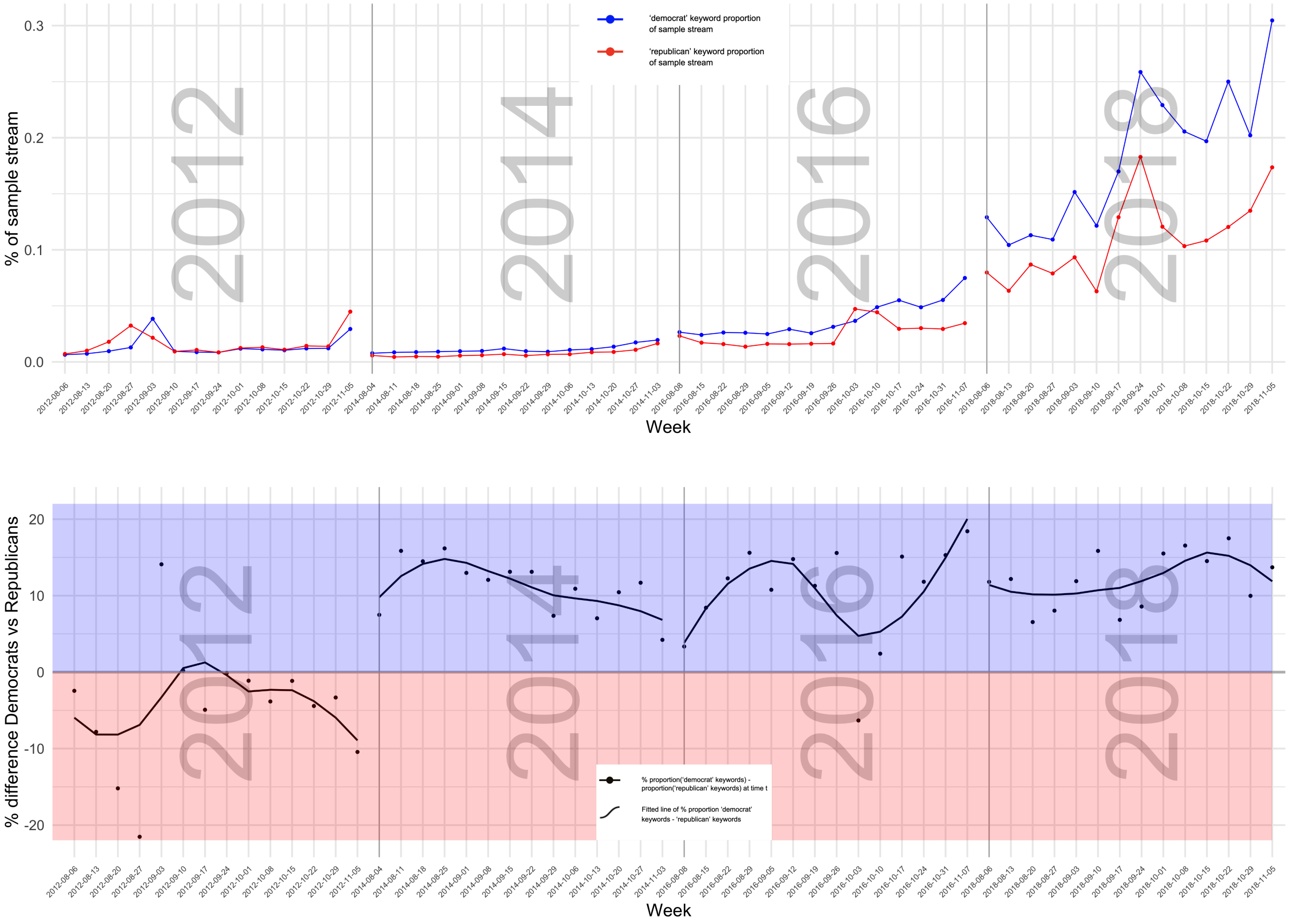


Figure 1.2: Mention percentages in sample stream; %"Republicans" - %"Democrats"

While it is clear that the terms "democrat" and "republican" have different meanings in other contexts besides U.S. politics - indeed, they are meaningful terms in their own right in most any English-speaking environ - these graphs illustrate some core issues with focusing on keyword frequency as the core quantity when using Twitter data (or, indeed, any text-based, user-generated social web data) as a measure for political/electoral support, and further emphasise Jungherr's (2015) argument shown above. Both graphs cover all data subset from the 1% sample stream, three months leading up to each country-wide election cycle (presidential and midterm) in the United States since 2012, totalling 4 elections, collapsed to the week-level. This specific choice of time range and keywords - both of which are reproducible over time, unlike candidate names - thus reduces the risk of unwittingly including large amounts of keyword mentions not relevant to the case at hand. The top graph depicts the percentage of the total sample stream, while the second, lower graph depicts the percentage by which the keyword "democrat" occurs more frequently than "republican" in the sample stream. According to this admittedly rudimentary 'forecast' - if the final week before a given election is understood as pertinent, Mitt Romney would have won in 2012 (incorrect), Democrats would have won the 2014 midterms (incorrect), Hillary Clinton would have won the 2016 presidential election (while she did win the popular vote by a considerable margin, she did not end up in the White House, so this forecast can be considered both correct *and* incorrect), and Democrats would have won the 2018 midterms in a landslide (correct). There also seems to be a clear trend towards a general 'advantage' - i.e. higher frequency - of mentions of the term "democrat*". While this can likely be explained in part by the demographic bias of Twitter's user base, it also highlights an inherent shortcoming of deriving any kind of meaningful information from keyword volume in tweets: merely mentioning something does *not* mean that the mentioner is also positively disposed toward it: if John Smith tweets "I hate the democrats", this adds to the total tally of "democrat*" volume, but, in the real world, is unlikely to translate to a vote for a Democratic candidate. While researchers have attempted to address this problem by analysing the *sentiment* of tweets in order to classify whether a given tweet mentioning a politically salient keyword is positive, negative or neutral (*sentiment analysis*), (see e.g. Bermingham and Smeaton, 2011; O'Connor et al., 2010; Ceron et al., 2014), there exists no convenient method for reliably classifying (political) tweet sentiment at scale. Furthermore, Figure 1.2 documents a noticeable rise in relevance for the respective topics being counted. While in 2012, tweets mentioning "democrat*" or "republican*" accounted for less than 0.1% of the Twitter stream, the same keywords made up up to 0.5% of the stream in any given week cumulatively in 2018[7]. In summary, it is clear that political tweeting has become more prevalent over time, and the pro-democratic bias of the Twitter population as a whole may be becoming increasingly more visible. But: this graph makes clear this particular approach to forecasting election results using volume metrics of keyword mentions on Twitter falls widely short of its goal - there are immeasurable numbers of known and unknown unknowns distorting outputs to the point where any degree of interpretation eventually descends into guesswork.

However, what if party names are exchanged for candidate names? Assuming that "democrat*" and "republican*" are too vague to capture specific Twitter activity related to American politics, perhaps opting instead for the names of those individuals on the ballot

[7] This trend occurred in the context of considerable overall growth in the number of tweets posted to the platform (InternetLiveStats.com, 2019)

paper as keywords to track allows for a more targeted and specific analysis of the merits of keyword mention volume as an indicator for electoral support. Figures 1.3 and 1.4 replicate the same week-by-week volume of Figure 2 for the candidates running for the US presidency, in 2012 and 2016 respectively - Barack Obama and Mitt Romney, Hillary Clinton and Donald Trump - for 3 months leading up to the election dates. In this case, both 'forecasts' correctly predict the eventual winner one week before the election date: Obama and Trump. This suggests that for a presidential election scenario, in these cases, when following keywords specifically referencing the candidates in a race, the mention volume approach performs well. Furthermore, the two sets of graphs (Figure 1.2 and Figure 1.3/ 1.4) seem to be showing some of the same trendlines: for instance, in the fifth week of the 2012 graphs, Obama *and* "democrat*" experiences a significant boost in volume. For 2016, the party and candidate graphs certainly do not align as well, but there is still a distinct similarity in the timing of weekly ups and downs. This suggests that the two distinct but conceptually very close keyword volume analyses show significant overlap, suggesting that - underneath the cloak of unquantifiable noise, something is being measured that pertains to public opinion, and thus electoral outcomes. Furthermore, unlike the majority of established polling and public opinion research output leading up to the 2016 US presidential election, this rudimentary candidate-keyword volume approach correctly predicts Trump as the eventual winner of the 2016 election, and further indicates a much tighter contest than was widely reported in the media and backed up by polling at the time. While this does not mean that this method is sufficient for forecasting election results, it certainly indicates that Twitter may contain signals relevant to public opinion which traditional polling is unable to capture, to a point where Twitter-based methods have the potential of outperforming established approaches. This thesis is devoted to producing evidence which can help illuminate the systematic ways in which these signals manifest themselves.

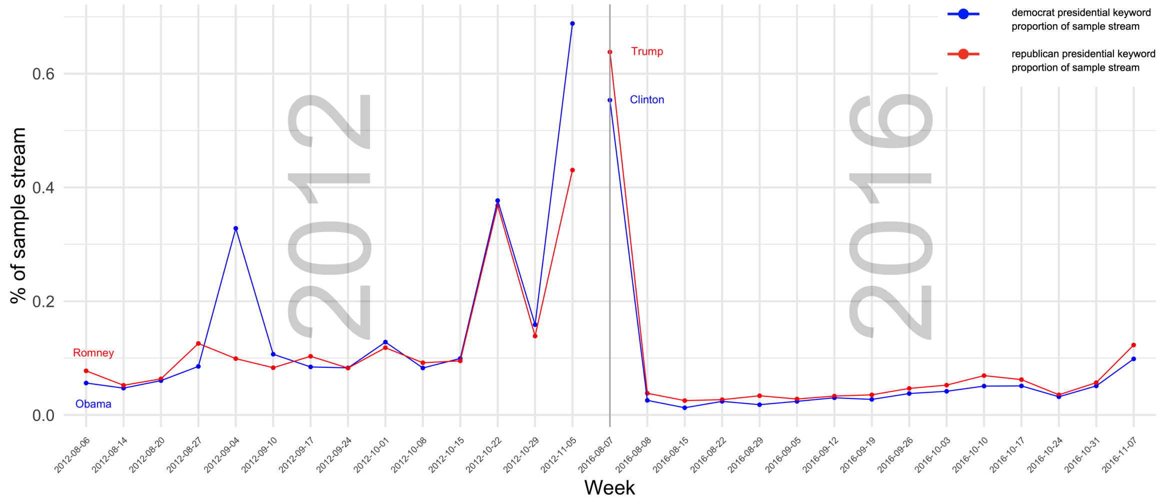


Figure 1.3: Percentage of mentions in sample stream for presidential candidates, 2012 & 2016

This thesis documents a comprehensive and in-depth investigation of what such signals mean in the electoral context. The key empirical and methodological contributions of this thesis are presented in three distinct and self-contained papers, written in a journal article format. This thesis format is advantageous for this research project, as it allows for a multi-faceted investigation of the problem space with different research strategies and underlying

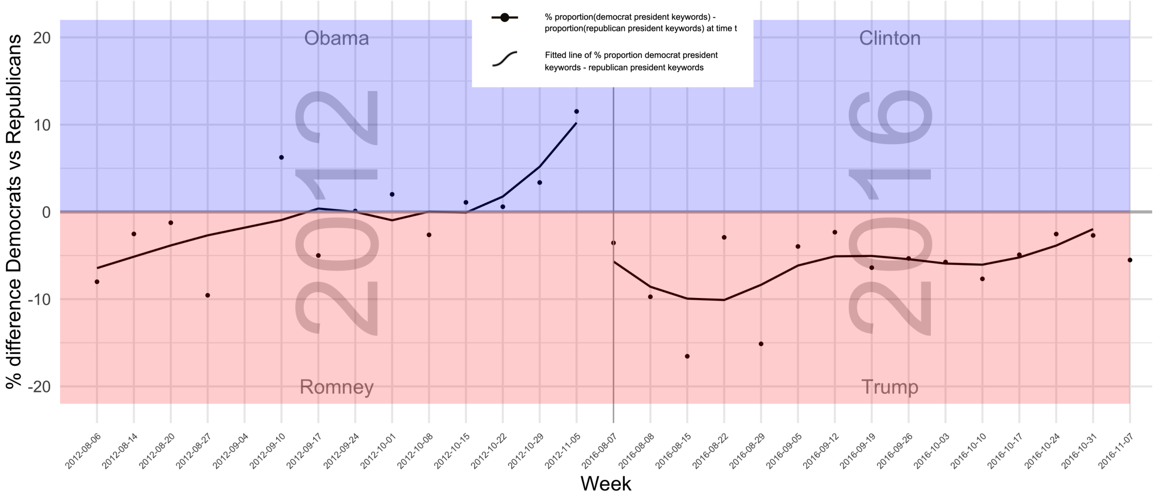


Figure 1.4: %Democrat candidate - %Republican candidate, 2012 & 2016

cases, with a focus on methodological experimentation and innovation and a broad array of applicable evidence, whereas a traditional book-chapter thesis approach would lend itself better to a singular, large-scale research study.

The first paper, 'Finding friends: a hybrid approach to geo-locating Twitter users' addresses the need for ways of classifying tweets and users with relevant demographic information given the sparse nature of Twitter's user profile. In this paper, I describe, validate and empirically apply a novel method for geo-locating Twitter users into politically relevant home locations at scale. These locations are categorised at five levels (municipality, administrative area, country, latitude, longitude). Using this inferred location information, Twitter users can be sampled (and, often more importantly, excluded from samples) based on their home location, so as to enable a more fine-grained and reliable study of relevant scientific questions using Twitter data. This is highly pertinent for questions of public opinion and voting, which differ greatly based on location. Besides the explanation and documentation offered within the paper, I also provide access to a software package, allowing researchers and any interested individuals to use it to apply my method for their own research purposes. Furthermore, the paper documents an example empirical application of this method in action, by investigating the geo-spatial dynamics of public opinion leading up to the Democratic presidential primaries of 2020. I find that using the geo-location method allows for a rich analysis of candidate support, which both mirrors and adds to conventional, poll-derived knowledge. This paper was co-authored by Jonathan Nagler (jn27@nyu.edu; conceptual, methodological and supervisory work), Andreu Casas (acs706@nyu.edu; coding, infrastructure, visualisations) and Nicole Baram (nhbaram@gmail.com; infrastructure). I performed the conception of the data pipeline and the empirical application, wrote the paper, performed the majority of coding, contributed to work on infrastructure and generated data visualisations.

The second paper, 'Understanding Political Sentiment: Using Twitter to map the 2016 Democratic primaries' tests previously used methods on new data, namely positive sentiment volume applied on Twitter data pertaining to the New Hampshire, South Carolina and Massachusetts Democratic presidential primary elections in 2016. Furthermore, this paper provides an exploration of a novel methodological approach of classifying the political sentiment of relevant tweet corpora and translating such sentiment distributions to electoral support, incorporating negative tweets into predictive models, a novel strategy in the literature. Twelve predictive models are applied to each of the three elections, showing that adjustment of samples by user-level home location improves the accuracy of established methods of sentiment analysis when the goal is extracting public opinion-salient signals from them. We further show that ordinal, intensity-focused machine-classification of tweet sentiment is a useful strategy for enhancing the degree to which political preferences can be extracted from electorally salient tweets. This paper was co-authored with Mark Elliot (mark.elliot@manchester.ac.uk). Conceptual and methodological work was conducted by Mark and myself, while coding, data visualisation, data analysis, data annotation and writing was done by me.

The third paper, 'Listening in on the noise: estimating individual-level political preferences from Twitter data' introduces a novel methodological approach to distilling individual-level political preferences from social web text using Machine Learning and distant supervision. I achieve this by filtering two distinct samples of Twitter users (one randomly selected and one selected purposively based on previously having shared election participation) for language clearly indicative of their voting intention, then manually label a given user who expressed such a voting intention with their likeliest vote choice. I then generalise the entire corpus of their tweets as a training set for a Naive Bayes classifier that predicts the likelihood of voting for party p based on the entirety of all tweets posted by all users in both samples. I apply this method to predicting popular vote share totals nationally and for California, Texas, Florida and New York within the 2018 US midterm elections, with differently specified models. Besides producing highly predictive results (smallest error margin: 0.15%), this paper empirically showcases the importance of sampling in Twitter-based political science research and allows for a conceptual analysis of the factors contributing to model performance. All work - conception, coding, data visualisation, writing - in this paper was done by me.

These three papers offer a wide-ranging and diverse contribution to the field. While all three papers focus on methodological innovation, ranging from new approaches to mapping sentiment analysis in tweets to aggregate vote shares, inferring individual-level demographic attributes and trialling new approaches of generalizing political stance and preferences to individuals based on their published online text content, all three papers also add theoretical foundations for the linkage between online political behaviour in the form of tweeting and observable offline political outcomes. This is particularly pronounced in my third paper, which tells a clear story of the relevance of certain topics to the 2018 election cycle, and how predictive they were of aggregate vote shares when aggregated up in-sample. Besides its immediate contributions, this thesis also provides a strong basis for the future examination of research questions about the nature of politics through Twitter data. For instance, by more accurately localising individuals geographically, it becomes possible to investigate questions of a politico-spatial nature through the lens of Twitter, such as whether the rise of anti-immigrant sentiment and support for radical right populist parties is linked with greater contact with immigrant populations or the reverse. Furthermore, by employing this thesis’ other methodological contributions, this for instance enables researchers to investigate local variation in individual-level policy preferences on key areas of public spending or development. It is therefore clear that this thesis has a strong potential of enhancing future political science research, as well as informing public policy.

Chapter 2

Related work

In this chapter, I discuss existing work relevant to this thesis. This literature review draws from a wide array of disciplines and fields, with the goal of illuminating the diverse topical threads which form the core investigative agenda of this research project. I begin by focusing on the field of public opinion research: how is it conceptually defined? Which theoretical frameworks have been employed in order to understand its formation? How can it be measured? What are contemporary critiques of public opinion measurement? Following this, I cover a further relevant field: forecasting and prediction, in both general and scientific contexts. I then introduce the political science literature concerned with forecasting election results. Finally, I zone in on social science research using social media data generally, and then focus on the key literature this thesis contributes to - Twitter-based election forecasting and public opinion measurement.

This thesis is to a large part concerned with investigating the efficacy and appropriateness of different methodological approaches for extracting public opinion and election-relevant signals from social web text. Hence, it is situated at the intersection of an emerging, interdisciplinary research methods agenda incorporating political science, data science and computational social science. Followingly, it is important to align and compare novel findings and approaches, which may on their surface be more closely related to fields such as computer and information science or linguistics, with established research in political science, which, while not necessarily addressing large-scale empirical questions in this area, this thesis primarily contributes to.

2.1 Public opinion research

2.1.1 An overview

Fundamentally, this research is concerned with ***Public Opinion*** - how it is to be understood, how it can be measured, and what it signifies, both in a descriptive and a normative sense. While this research instrumentally centres in on elections - a regular, quantifiable manifestation of what contemporary parlance intuitively understands as 'public opinion' - it is useful to understand the origins and development of the term, and how it can be delineated from other concepts and constructs. Encyclopædia Britannica defines public opinion as "an aggregate of the individual views, attitudes, and beliefs about a particular topic, expressed by a significant proportion of a community" (Davison, 2017). This definition gets at the core facets of public opinion as we understand them today: it is an *aggregation* of individual-level viewpoints, for a "community" - this could be as large as the population of a country, a professionally, ethnically or otherwise defined group, or as small as the residents of a street. Furthermore, it captures the notion that public opinion is not necessarily a monolithic consensus, but rather the sum of individual views, and thus typically diverse, broad and subject to change. However, it is important to note that public opinion means different things in different contexts, as it is one of the most relevant concepts for the entirety of the social sciences - historians and psychologists may have different definitions of the term from political scientists, for whom it is relevant as "denominat[ing] the relationship between the government and the people" (Donsbach and Traugott, 2007, p. 2). In this thesis, the fundamental focus however is on what public opinion means for political science.

The dynamic nature of public opinion is clarified by the sociologist Charles Horton Cooley, whose 1918 book "Social Process" provides early insights into the formation of public opinion: "Public opinion, if we wish to see it as it is, should be regarded as an organic process, and not merely as a state of agreement about some question of the day. It is, in truth, a complex growth, always continuous with the past, never becoming simple, and only partly unified from time to time for the sake of definite action." (Cooley, 1918, p. 378). The German sociologist Ferdinand Tönnies delineates public opinion as being to society what religion is to community in his seminal 1887 (re-issued 1912) book "Gemeinschaft und Gesellschaft" ("Community and Society"). Both are shaped by individuals with faith and "doctrin" - i.e. views/opinions/beliefs - to form religion and public opinion (öffentliche Meinung) at the aggregate level. He argues that while religion rules over ("stellt sich über") the community ("Gemeinwesen"), public opinion rules over matters of the state. In practice this means that public opinion will judge political output as good or bad. However, Tönnies understands the process of even having "doctrines" necessary for the formation of public opinion as complicated, and reserved for few intellectual and wise people. He further indicates that contributing to public opinion formation requires motivation to understand the concepts and feed back into the political system (Tönnies, 1912, p. 268-272). Clearly, this understanding of public opinion, especially when conceptualised as parallel to religion, is one of elite political participation in the proto-democratic Bismarck era of Germany, and perhaps not entirely applicable to modern-day society.

The role of elites in public opinion is nonetheless highly relevant throughout the literature on public opinion, such as the eponymous 1922 book by Walter Lippman. He highlights the importance of individual-level perceptions of the world in shaping views, beliefs and opinions, and how perception invariably results in distortion and simplification. According to Lippmann (1922), humans construct their own "pseudo-environments" (p. 25), which they use to ascribe meaning to the *real* environment, which however in its vastness escapes the limits of human comprehension. Crucially, he argues that public opinion is shaped primarily by the media and news reporting, which in turn is made up by humans, whose reporting and writing originates, by definition, from their individual pseudo-environments. This, then, provides society as a whole with a constrained range of views and opinions, contingent on the pseudo-environments of members of the (media) elite. The problematic nature of the news media as the key shaper of public opinion is further exemplified by structural factors which determine what gets reported, how, and when. Specifically, the industry's focus on *novelty* (as opposed to continuing issues), reporting on *events* (rather than e.g. societal or cultural processes), and focusing on high-profile *leaders*, i.e. prominent and important individuals in place of a broader representation of all of society's groups and institutions, results in a distorted range of information from which the public can derive their views, attitudes and beliefs, thus considerably constraining the forms public opinion can take. This leads to an imbalance of power in the relationship between the state and public opinion: "Yet democracies, if we are to judge by the oldest and most powerful of them, have made a mystery out of public opinion. There have been skilled organizers of opinion who understood the mystery well enough to create majorities on election day." (Lippmann, 1922, p. 254). This further corroborates Cooley's (1918) argument of the unification of public opinion for specific action.

When translating these early conceptualisations of public opinion into the modern, digital world, it intuitively appears that some factors hold true while others have changed given the seismic shift in how information flows. If every individual with a social media account now has the potential ability to be a producer of 'content', and this content has the potential of influencing individuals' views, this would suggest that public opinion now has the potential of taking a much broader range of forms. Further, if the "pseudo-environments" paradigm holds true, it may arguably be the case that in the context of social media, the content creator's framing of the world still has the potential of constraining the range of views, attitudes and preferences of their consumers. I argue that this conjecture can by hypothetically supported by the social media-driven emergence of new far-right movements in the form of "gamergate", "Qanon" and the "Alt-right" - previously taboo positions emerged by being pushed by individuals not affiliated with the mass media, promoting a certain highly constrained narrative of what the core issues facing the world at that given time were supposed to be. And, as suggested by Cooley, this 'new' ideology permeated into the mainstream, meaning that it must have driven many individuals to change their mind in its favour.

If this is indeed the process by which individuals acquire salient views and opinions, the questions of how exactly information 'trickles down' from elite discourse to 'the woman on the street', and which organising principles and/or heuristics she uses to process this information into pertinent views, remain. Converse (2006) (originally published in 1964) provides a good starting point for understanding this process. This groundbreaking paper seeks to understand the organizing principles individuals employ when shaping their own political views and attitudes. In short, Converse finds that the majority of individuals do *not* have an internally coherent set of political views and attitudes, and do not form them based on what may be understood as ideology. Rather, the 'woman on the street' picks and chooses the current political message distributed by the mass media and political elites on a given day, and forgoes the step of testing if this is compatible with her established belief system before integrating this position into it. In other words: the average person is inconsistent in how they ideologically organise their acquired political knowledge, if they do so at all. Hence, public opinion, as communicated through survey-based findings, should not be understood as conveying a widespread comprehensive understanding of what the issues of the day mean in the broader context of politics. Further, it should not be understood that the majority of those individuals who were surveyed have any long-lasting attachment to the specific answers they gave. The degree to which this is the case is highly contingent on individuals' levels of education and their overall socio-economic status. And, indeed, Converse's assessment of individuals' ideological malleability holds true in the anecdotal social media / alt-right example: if someone is able to reach a receptive audience with appropriate messaging, they are likely to be persuadable, oftentimes in spite of previously held positions inconsistent with these new viewpoints.

This skepticism of viewing polls and surveys as an accurate representation of ’true’ public opinion is shared by a further influential scholar of public opinion formation, John Zaller. In his 1992 book "The Nature and origins of mass opinion", he goes as far as to state that "most of what gets measured as public opinion does not exist except in the presence of a pollster" (Zaller, 1992, p. 265). This assessment illuminates why exploring non-survey based approaches to measuring public opinion - such as those derived from Twitter - may be fruitful. If individuals only develop attitudes (or, worse yet, *pretend* to do so) when they are asked about them, they perhaps shouldn’t be treated as such. Indeed, an approach that relies on observational rather than intentional data may provide a framework for a more realistic reflection of public opinion.

Zaller’s work also contributes significantly to the understanding of individual-level political attitude formation, building heavily on Converse’s work by formalising an empirically testable theory of how information flows from elites to average individuals and thus shapes public opinion. In his "receive-accept-sample" model of individual information acquisition, Zaller emphasises the importance of what he refers to as "considerations", a "compound of cognition and affect - that is, a belief concerning an object and an evaluation of the belief" (p. 40). Zaller argues that considerations result from political messages originating from elite political actors and typically disseminated by the mass media[1] which, through interpretation on the side of the individual, form a repertoire of considerations they can fall back upon when evaluating new messages and information. Building on Converse’s (2006) findings, Zaller defines four ways in which individuals vary in how they acquire and process considerations. First, individuals vary in how likely they are to be exposed to such messages, and how likely they are to extract information necessary for constructing considerations from them - factors determined by relative levels of "political attentiveness" and "political awareness" (Zaller, 1992, p. 43). Second, people differ in how likely they are to "resist" new information which contradicts previously formed considerations, which may - if only in the individual’s mind - be consistent with their internal logic. Again, the author suggests that political attentiveness is key (p. 44). Third, Zaller talks about the "accessibility axiom". Here, it is less dependent on individuals’ political awareness, but rather on the time since a relevant, related consideration was recalled. As previous considerations need to be cognitively activated in order to contextualise incoming information, the time since a previous consideration was made is highly relevant for the processing of a new message (p. 48). Fourth, and this ties into how individual-level views and attitudes feed into and shape public opinion, Zaller notes the "response axiom". This means that individuals will respond to survey questions by selectively retrieving considerations they have made, and potentially with different frames in mind. In other words, "individuals answer survey questions by averaging across the considerations which are immediately salient or available to them" (p. 49).

Both Converse and Zaller’s work highlights the complexity of the term ’public opinion’ - at every stage, it is contingent on both the individuals that make up a given ’public’, but also on those that measure it. Given that most individuals do not seem to care that much

[1] In the modern social media landscape, it seems that the importance of the mass media as a vehicle for political messages is no longer essential

and come up with their answers to pollsters *on the spot*, this suggests that public opinion is open to manipulation from the political and media elites that steer it. In contrast to this view of an "ignorant public" stands that of Lupia et al. (1998), who argue that individuals (voters) do not *need* to have full information and well-thought out opinions on the intricacies of politics. Rather, they are able to make reasoned choices given the system of cues and shortcuts that is provided to them by trusted sources in their immediate network, as well as built-in safeguards in the institutional make-up of the political system. The authors use the analogy of traffic signalling to illustrate their core argument: "Advocates of complete information might argue that successful automotive navigation requires as much information as you can gather [...]. At many intersections, however, there is a simple substitute for all of this information - a traffic signal" (p. 7).

While arguably addressing different aspects of the conceptualisation of public opinion, I would suggest that Zaller's framing of the issue has aged considerably better than Lupia's. Given the current context of widespread online disinformation campaigns (see e.g. Freelon and Wells, 2020) and a rise of far-right parties across seemingly mature democracies (see e.g. Rodrik, 2020), it would appear that in some cases, traffic signals are not sufficient for preventing traffic deaths - for instance if the technology tasked with its proper operation is no longer being maintained (consider, for instance, the decline in local news reporting across western countries, and the dearth of available essential information that follows from it), or if the driver is intoxicated (such as in the alt-right example, where the 'opinion leader' is someone on the internet whom you have never met but choose to trust because you have no friends in real life).

In political science research, as well as in general conversation in democratic regimes, public opinion becomes most salient when it comes to its electoral implications. A further canonical work investigating this facet of public opinion is Lazarsfeld et al. (1944). By following a panel of Ohio-based voters over the course of the 1940 US presidential election campaign, the authors illuminate several processes and patterns of electoral public opinion which are seen as common knowledge in the present day. For instance, the authors identify clear patterns between individuals' socio-demographics and both their *a priori* stated voting intention and their eventual vote choice, allowing them to confidently predict eventual vote choice for initially undecided participants. Furthermore, the authors find that the vast majority of participants, and, by extension, voters, do not change their mind regarding the party they want to vote for over time; whereas a sizeable minority - the "changers", do. These changers are further sub-divided into different archetypes: the "crystallizers", who go from "don't know" to a vote choice, the "waverers", who change their minds mid-campaign and eventually return to their initial preference and the "party changers", who voted for a party other than their declared preference at *t1*. While such changers are far less common than those who stick with their party preference across election cycles, and indeed pass it on over generations, their relevance to understanding variation of electoral outcomes is disproportionately large, especially in a two-party system like the USA's. When comparing this to modern-day electoral politics and electioneering, it explains in part why parties target specific swing states and counties with disproportionate resources for political campaigning (i.e. an attempt to persuade changers of the merits of their party/candidate), whereas electoral outcomes in most other geographic areas are treated as de facto pre-determined.

Beyond (electoral) politics, public opinion is an important concept for other aspects of the world, be they the private sector or non-governmental organisations (NGOs), or matters of state. For actors in the private sector who seek to make a profit by selling a service or product, positive public opinion of their company or brand is essential. For matters of state beyond elections, the question of how important (positive) public opinion is to continuity and stability is a more complex one. This has been studied especially in the context of foreign policy and foreign military involvement. Early scholarship on the subject resulted in the so-called *Almond-Lippman consensus*, which argued that voters cared little about foreign policy, war, and so on, and thus states had more or less free reign to conduct any kind of foreign policy without fearing electoral reprisals (see e.g. Holsti, 1992). However, following the Vietnam war and the widespread mobilisation of anti-war sentiment throughout the USA, as well as more recent strong opposition to the US's continued presence in Iraq and Afghanistan, it appears that favourable public opinion is indeed required in order for a government to conduct their desired foreign policy, at least when this includes military intervention.

In conclusion to this section, I have demonstrated that public opinion is vast, complex, dynamic, differently relevant in different arenas, and subject to being influenced, or even manipulated. The media plays a large role in shaping public opinion, and at the same time political actors, especially those holding elected office or seeking it, are highly vulnerable to it while simultaneously being dependent on it - even in proto- or non-democratic societies - for the successful implementation of their desired agendas, or for achieving (re-)election. Opinion among scholars is divided as to whether average citizens have any substantial amount of pertinent knowledge from which to develop consistent political views, and further, whether it is even important for them to do so. I argue that the crucially important reason *why* it is important to study and understand public opinion is the normative value of responsiveness which democratic regimes are often held to. Without an understanding of what it is that citizens need and want, governments and bureaucracies' jobs will consist of - at best - guesswork, or - at worst - governance in the interest of privileged elites. Hence, it is essential for a society to have a broad understanding of what it is that the public thinks and wants, in order to organise long- and short-term agendas of the state accordingly.

The mechanisms by which public opinion is shaped that I have discussed intuitively hold true when transposed to the social media era. However, I argue that the role of the mass media as a gatekeeper of which individual-level positions and political views can feasibly contribute to public opinion has waned significantly in the modern day, leading public opinion to become more multi-faceted. Further, the understanding of Zaller and Converse that much of what is commonly referred to as public opinion has only been measured because pollsters primed and framed respondents in a way that suited their questioning agenda, suggests that a conceptualisation of measuring public opinion using observation rather than, or as well as, intervention - such as is proposed in this thesis - may be a fruitful contribution toward a more accurate measurement of the latent variable that is *true* public opinion.

2.2 Survey research: an overview

In the following paragraphs, I begin by outlining the history of survey research, followed by some basics of survey design. I then discuss the criticisms of survey research and outline alternative approaches, with a particular focus on using observational data, such as official government documents, or *tweets*.

A brief history of survey research

The quintessential instrument for measuring public opinion[2], is the *survey*. In essence, a survey is a formal question-answer routine: multiple respondents are asked the same questions and typically provided with response options, their answers are aggregated, and insights are derived therefrom. The details matter greatly in terms of how useful the output from a survey is. For instance, *who* gets asked in the first place, i.e. recruitment of participants, and prior to this step, the definition of a sampling frame, i.e. the pool of characteristics which define the population a survey is studying, have a large impact on the output of any given survey. Furthermore, details such as the specific wording of questions, or the number and type of provided response categories can result in widely different results.

The turning point toward widespread adoption of surveys for fact-finding for business and government alike came from the discovery that probability sampling allowed for "bias-free estimates and measurable sampling errors" (Groves, 2011, p. 862). This innovation, widely attributed to Polish-American statistician Jerzy Neyman (1934), but built on work by Kiaer, Bowley and Fisher (see e.g. Converse, 2017), argues that a cost-effective and reliable approach to obtaining insights on large, diverse populations lies in random probability sampling of said populations with a focus on representativeness; this way providing every element of the population with the same probability of being included in the sample. From hereon starts what Groves (2011) refers to as the first era of survey research, dominated by face-to-face interviews conducted by large survey research firms, such as Gallup, and the US federal government. Groves notes that this was also the golden era of respondent participation: "With response rates often over 90 percent, there was more concern with contact rates than refusal rates" (Groves, 2011, p. 863). This era also sees some significant conceptual and methodological developments in the field of survey research. While initial "social surveys" had broadly sought to ask for objective metrics - 'How old are you?', 'How many people live in this household?', 'What is your annual income?' - the mid-20th century saw a shift toward asking about people's views and opinions, their attitudes and domain-specific knowledge. This development was pioneered especially by the aforementioned Gallup company, but also by contemporaries in academic social science, such as Elmo Roper and Paul Lazarsfeld. Moreover, with the introduction of the *Likert scale* (Likert, 1932), an instrument that provides respondents with multiple quasi-sliding scale answer options to a statement, such as "agree completely", "agree somewhat", neither agree nor disagree", "disagree somewhat" and "disagree completely", a very cost-effective, and quantifiably reliable[3] method for collecting and measuring subjective data. With this innovation, and the

[2] And indeed for creating the vast majority of empirical data used in most any piece of social science research

[3] compared to the established approach at the time, *Thurstone scales*, which necessitated interviewees to mark statements they agreed with, the sum total of agreed with statements would then convert into an aggregate score for the desired measure

understanding that it allowed questioners to find out what people think about *n*, the sample survey rose to dominance beyond academic research and government, and took root as a staple in the corporate world, as it allowed businesses to find out what customers wanted (market research), but also to track how their brand/product was perceived. Hence, the sample survey has inarguably contributed greatly to the development and growth of the marketing, advertising and consulting industries, as well the consolidation of consumer capitalism as the dominant economic and cultural paradigm of the 20th and 21st centuries.

Throughout this period of growth in the survey industry, it became clear that asking people the same question repeatedly over time is a very useful tool that allows for comparability and the identification and assessment of trends. This is particularly important for survey-derived public opinion: by asking the same questions over time, it became possible to dynamically follow the 'mood of the nation', and adjust policies and electoral campaigns accordingly. Examples for such recurring surveys are typically found in large survey research companies and government-backed/run surveys, such as the American National Election Study (ANES) which has questions that remain the same since 1948, the British election study (since 1964), the Eurobarometer, a European Union-wide battery of questions on e.g. social attitudes and the economy (since 1973); or, for private / non-profit survey institutes, such as Gallup, Roper or Allensbach, who provide access to survey data as old as 70 years on their websites, where several questions of continued importance - e.g. regarding the economy - are comparable over time.

A particular type of survey is the *poll* (often prefaced with 'opinion'). Deriving its name from the term for casting a ballot in an election - '78% of people went to the polls on election day', the modern opinion poll typically asks a probability sample of individuals a politically/electorally pertinent question aimed at eliciting a subjective data point from the respondent, such as: 'Who would you vote for in the upcoming election?', 'On a scale from 0 to 10, how do you rate President Trump's management of the economy?'. Polls came to prominence in what Groves (2011) describes as the second era of survey research ("The era of expansion"), aided by the widespread adoption of landline telephones in the vast majority of households in industrialised countries (p. 865). By randomising the non-area-specific digits of phone numbers, pollsters were able to easily obtain probability samples of target populations of respondents. Later yet, the introduction of Computer-assisted telephone interviews (CATI) - at times not even assisted, but computer-led, and later yet, the introduction and evolution of internet-based polls further fuelled the expansion of the polling industry.

Opinion polling in 2021

At the time of writing this thesis (2021), the average citizen of a developed country - most prominently so in the USA - finds themselves inundated with reporting on the findings of opinion polls most any day of the week. Such polls vary widely in what they report, *even if they are asking precisely the same questions*, of the same population, in the same time frame. For instance, for the single day of August 21st, 2019, the poll aggregation section of fivethirtyeight.com documented the publication of 17 different polls, conducted by six different polling companies, solely on the topic of the 2020 US Presidential election (see Appendix C.1 for screenshot) and Presidential approval. The reported polling findings diverge significantly and often beyond mere sampling error - for instance, in the provided example, disapproval towards President Trump ranges from +6 to +14 - a staggering 8 percentage points.

Despite this quasi-inundation with poll-based empirical evidence, we currently find ourselves in the third era of survey research (and polling) (Groves, 2011) - one of pronounced, and continued decline in respondent-level engagement and ever-sinking response rates to surveys. Hence, survey researchers and pollsters have developed an ever-more refined toolkit of approaches to deal with non-responses, such as over-sampling of certain demographic groups followed by advanced statistical approaches to de-biasing samples and making them representative of target populations. Furthermore, the number of surveys and opinion polls now conducted online, through large-scale pre-recruited panels of supposedly representative individuals, has risen significantly, and seems to be the industry's main approach to dealing with the rising cost of obtaining a reliable representative sample of respondents in light of very low telephone survey response rates. Furthermore, the transition away from landline to mobile telephones no longer allows for confident allocation of randomly selected phone numbers to a narrowly defined geographic entity, further lowering the *a priori* reliability of representativeness when using the established methodology of sampling.

As stated in the introductory chapter, the polling industry has been subject to widespread critique in the recent past, in light of prominent inaccurate immediate pre-election polls of high-profile elections, including the 2016 US presidential election, the 2015 and 2017 UK general elections, and the 2016 UK referendum to leave the European Union. While I do not wish to speculate as to the true reasons behind these failings, it does appear that several factors are culminating in making accurate polling harder in the modern day. First, the aforementioned decline in survey response rates and landline telephone adoption has something to do with this. Second, it is worth re-iterating Zaller's assessment that many individual-level considerations do not exist outside of the minds of pollsters. While this may not necessarily be pertinent for simple vote choice polling, it clearly gets at an issue whereby survey respondents are faced with numerous options, none of which reflect their actual preferences. Most importantly however, I argue that it is also worth considering factors such as social desirability bias: respondents are "afraid" to share supposedly unpopular opinions with pollsters but have no qualms to do so at the ballot box. This idea is supported in the case of the 2016 UK Brexit referendum by the fact that online polls consistently indicated a victory for the leave side, while phone polls showed remain ahead (Clark, 2016). This is supported by empirical evidence indicating that phone survey responses are more prone to responses shaped by social desirability bias than internet surveys, which are, by definition, unsupervised (Christian et al., 2008).

2.2.1 Survey design

A well-theorised and specified design is an essential component of a reliably and validly working survey instrument. As such, it is necessary to briefly outline key problem areas typically taken into consideration when designing surveys, as well as particular aspects of survey design. Rather than providing a guide to designing surveys, I seek to illustrate the complexity of survey design and its ability, when badly executed, of biasing any insights it aims to facilitate. Furthermore, as real-world social research usually operates within tight resource constraints, this becomes even more pertinent when attempting to objectively evaluate the usefulness of surveys for public opinion data collection.

The survey design process sits broadly in the middle of the chronology of the research process (Punch, 2003, p. 8) - it requires the researcher to have already established their broader area of research, defined workable research questions, and come to the conclusion that answers to these questions would best be found by means of survey research[4] - i.e. asking several *individuals* a number of questions aimed at measuring the core concepts at the heart of the research question. In most cases, survey questions would aim to produce data that can be analysed *quantitatively*, by providing respondents with a pre-specified set of possible response categories instead of allowing free-form answers to questions. Punch (2003) describes the core of the *quantitative research strategy* as follows: "Assuming [...] the individual person is the unit of analysis, the essential idea of the quantitative survey is then to measure a group of people on the variables of interest and to see how those variables are related to each other across the sample studied" (p.23). Beyond the specification of precise research questions, this means that the following broad survey design steps are to be addressed by the researcher:

1. What are the **variables** that can answer the research question? How can these variables be operationalised, if necessary, in order to represent measurable constructs? What data are required in order to provide answers to the research question using these variables? How to ask the questions sought to provide pertinent data on variables of interest (e.g. wording, timing)? How to deal with non-attitudes?

2. What is the population this research aims to learn more about? How can a **sampling** frame be decided? Which sampling strategy will be applied given resource constraints?

3. **How will the survey be delivered**? Face-to-face interviews, telephone interviews, mail-out questionnaires for self-administration or online questionnaires?

Unsurprisingly, there are several possible pathways for addressing the three above points. However, different choices will produce different survey outcomes and thus different research findings. In reference to **variables, question wording and operationalisation**, this is most easily shown by focusing on the wording of questions, and the way in which seemingly subtle differences in wording can produce widely different data, both within the same survey and the same sample, and across both different applications of the same survey and different surveys. Real-world evidence of this phenomenon is provided in Iarossi (2006): "A Latinobarometro poll in 2004 showed that while a clear majority (63 percent) in Latin America would never support a military government, 55 percent would not mind a nondemocratic government if it solved economic problems" (p.1). The author reports this effect across several countries in which the question was asked (p.2), in order to underscore their point that "the way a question is worded [can affect how respondents answer it], in order of up to 30 percent change in attitude". While stating that there is no one-size-fits-all guide for question wording, Iarossi (2006) defines three basic principles to keep in mind when formulating survey questions, ***relevance***, ***accuracy*** and ***willingness***. The author suggests 'relevance' is achieved when "the questionnaire designer is intimately familiar with the questions" (p. 27), meaning that the linkage between what is aimed to be measured and

[4] rather than other existing methods of empirical data collection in social science, such as interviews, focus groups, participant observation, experiments, document/official record analysis, discourse analysis, to name but a few ...

what the questions actually measure needs to be well-understood and implemented in the question wording. In regard to 'accuracy', Iarossi explains that it is achieved in the survey design context, if "a question [...] collects the information sought in a reliable and valid manner" (p. 27). Hence, it is to be understood an amalgamation of the scientific concepts of reliability and validity of a measurement instrument. Finally, 'willingness' describes the survey design task of implicitly encouraging respondents to participate in a given survey, and further to maximise the length and breadth of said participation (p. 28).

A further crucial step of survey design is **sampling**. As stated in the previous section, the gold standard for conducting high-quality research in the social sciences assumes the existence of a randomly selected probability sample of a pre-defined population of individuals[5]. This approach gives every element of the population the same probability of being included in the sample, and thus, if the sample is large enough in size, will closely approximate the characteristics of the population. This process is typically referred to as simple random sampling. Sampling is a complex statistical process that deserves a large amount of attention which I cannot cover in detail in this section. For more on the principles of probability sampling for surveys, see e.g. Marsden and Wright (2010). Instead, I discuss the real-life difficulties of obtaining such a gold standard sample for the average social science researcher. A common resource-conscious approach to sampling is *stratified sampling*, whereby certain characteristics (strata) of a population are known, allowing the population to be divided so as to draw samples from each of these strata. Such strata could be, for instance, membership in a racial/ethnic group. However, in many cases, researchers find themselves presented with an imperfect sample for which survey data has already been collected, leaving them with *post-stratification* as the standard approach to making a sample more representative of a target population. Often, this is achieved by employing *weights* for certain strata. Using the example from above, this would mean adjusting the number of African-American respondents in line with their proportion of the target population. Weighting also becomes highly relevant when dealing with non-response in surveys.

When considering **the method of delivery** of a survey questionnaire, researchers are faced with a choice between face-to-face interviews administered by professionally trained interviewers, phone interviews conducted by professionally trained interviewers, phone interviews conducted by automated computer systems, mail-out questionnaires to be completed by respondents, and online questionnaires, to be completed by respondents. The cost associated with these approaches declines more or less in the order mentioned, but the essential question is - will the choice of survey delivery type affect substantive findings derived from a survey? There is a considerable amount of evidence to suggest that it does. First off, there is a body of work that suggests that a crucial factor, regardless of whether a survey interview is delivered face to face or on the phone is whether a human leads the interview. If an interviewer is present, this typically predicts lower item non-response than if they are not, but higher non-response overall (Hox and De Leeuw, 1994). Furthermore, respondents are significantly more likely to divulge sensitive information in survey delivery modes without an interviewer (Hox and De Leeuw, 1994; Tourangeau and Smith, 1996). When comparing online survey data to traditional forms, there are striking differences in the data they generate. Duffy et al. (2005) find that online panellists (recruited, as described above, in

[5] Surveys can also be addressed at non-individuals, such as businesses, government agencies, NGOs, and other organisations/institutions, but for the purpose of this thesis, I limit my analysis to surveys addressed at individual human beings.

a non-probability sampling framework) differ significantly from traditional samples: they are more politically active and more likely to vote - and in other words, not representative of target populations, even after application of post-stratification methods such as propensity score matching. These findings are mirrored by more recent research by PEW, which finds statistically significant differences in responses to questions ranging from 2 to 18 percentage points between online and phone surveys (Keeter et al., 2015).

In this overview of different problem areas researchers concerned with designing surveys are faced with, I have outlined the immense complexity of the survey design process. While there are clear recommendations for best practice, many such recommendations are difficult to fulfil given the real-world constraints social scientists, and indeed for-profit pollsters, are faced with when conducting their research. Furthermore, on many occasions, there is no clear recommendation that can be made. For instance, more complex yet precise question wording may come at the cost of increased item non-response, whereas a more easily understandable question may increase response rates at the cost of an instrument that does not reliably measure the variable it seeks to measure. Clearly, survey design is a complex matter to which many researchers have devoted their entire career, and the current situation of declining response rates and self-selected online panels makes their work all the more important and challenging.

2.2.2 Critiques of the survey paradigm and alternative approaches

Having established that a) surveys are used as the go-to method for most empirical social science research and b) are absolutely ubiquitous when it comes to public opinion research, the question of whether surveys are the best tool available for generating empirical social data, and further, in which specific scenarios this may or not be the case requires further attention. This question goes beyond the contemporary issues associated with declining response rates and landline telephone connections (i.e. issues associated with the feasibility of conducting surveys, not the quality of the resulting data), but rather goes to the root of whether the paradigm of treating individuals' answers to questions is reliable data.

Fundamentally, research findings derived from surveys treat individuals' answers to questions as factually true information [6]. However, I argue that to some degree, a statistical approach to modelling what is essentially *lying, forgetting/mis-remembering* or *lacking knowledge coupled with embarrassment* seems somewhat misguided, as I would argue there is an essential human-emotional quality at the core of at least some reasons why respondents don't provide the 'true' answer. This may not be easily modelled, and, if it can be, is likely to be validated with yet more survey data. This completely forgoes the matter of whether they should be. We know, for instance, that some survey respondents - like any human - lie. This is more pronounced when it comes to personally sensitive questions (see e.g. Harrison, 1997; Wyner, 1980; Hancock et al., 2007) than seemingly trivial matters like one's preferred breakfast cereal, but is nonetheless present anywhere where humans answer questions (see e.g. Guess et al., 2018). This core issue with survey data was outlined in the expansionary period of survey research by sociologist Thomas Rhys Williams in the very first issue of Public Opinion Quarterly:

[6] With the noteable exception of the widely studied phenomenon of measurement error in surveys, which e.g. identifies false information in responses as correlated to non-response bias, and thus not randomly distributed among survey respondents (see e.g. Tourangeau, 2003; Olson, 2006)

> *The assumption that a reply by a respondent to a question is "the answer" insofar as his social behavior is concerned is fallacious. This contention is by no means novel; Linton, La Piere, Merton and Deming, among many others, have pointed out that in most societies there exist institutionalized patterns of saying one thing and acting to the contrary* (Williams, 1959, p. 57)

However, this only gets at an aspect of surveys where respondents choose not to reveal something, or wish to seem as if they had knowledge necessary for answering a given question when in fact they do not. There is a greater issue with survey-based data which stems from individuals needing to recall past behaviour, attitudes and beliefs. While this may arguably be easier for questions such as 'did you vote in the last presidential election?' than questions like 'what did you have for lunch two weeks ago?', there remains the issue that individuals often do not recall autobiographical episodes, however relevant any researcher thinks they may or may not be (see e.g. Bradburn et al., 1987). This has the potential of causing individuals to misconstrue their responses - both intentionally and not.

Besides approaching the issue of insufficient reliability of surveys as a problem of survey design, or of respondents' willingness or ability to render a factual response, there is a large-scale literature devoted to the issue of measurement error more generally in surveys. Bound et al. (2001) investigate the claim that measurement error in a given variable will be random, and thus not related to the 'true' quantity of said variable - finding that this does not empirically hold true. Beyond this, there is a large body of literature which empirically investigates the determinants, and more importantly impacts and biases of such measurement error, e.g. Hyslop and Imbens (2001); Kapteyn and Ypma (2007); Schennach (2016). It is important to note the advances made throughout this literature, regarding the ability to correct both statistically and conceptually for measurement error.

Beyond the factors that respondents may lie and may have unreliable memories, there remains the core issue that surveys form an *intervention* resulting in *non-observational* data. In other words, surveys produce *intentional* data which would not have been generated otherwise. This is in contrast to *consequential* data, which are generated by individual behaviour (often online, or mediated by technology, such as credit card transactions), but often contain ample information that can be mined from them, e.g. by researchers (Purdam and Elliot, 2015). When considering the evidence presented in the above section about public opinion and relating it to the outlined issues that exist with the survey paradigm, it followes that respondents may not a) be interested in, or have any knowledge on the topics a survey is measuring and b) may construct attitudes on the fly while c) potentially being unable to express their *actual* views and attitudes on the same or related topics because the closed-ended format of most surveys does not allow for it. So, assuming that a number of respondents do not have an opinion toward *n* prior to participating in a survey, is the datum documenting their *ad-hoc* opinion useful to furthering the understanding of what a certain target population thinks/knows about *n*? While one could argue that non-attitudes are indeed attitudes, this would require confidence that respondents actually report non-attitudes as such instead of picking other available, value-laden response options. While proponents of surveys would argue that one could easily integrate knowledge testing questions into the survey design, or repeated questions aimed at measuring the respondent's attention and involvement, so as to discard responses deemed as 'invalid', this may in turn negatively affect overall response rates. In any case, while this may be a useful tool for improving survey data validity, it misses the core problem associated with non-observational data: *everything* here is 'made up' in the sense that it only exists because someone asked in the first place. But, given the ubiquity of survey data across the social sciences, there is little that they can be benchmarked to[7] - surveys are the de facto 'gold standard'.

I suggest that studying public opinion - and, indeed, other phenomena - using digital trace data has the potential of complementing and rivalling surveys in several regards. Due to their nature as observational data, the crucial element of interviewer intervention; of *asking*, is not present. Rather, the individual-level behavioural datum exists because their author *wanted it to exist*. I argue that this may be particularly useful in identifying what the broad parameters of public opinion *are* in the first place. To return to the anecdotal example of the emergence of the alt-right: mining Twitter data would provide researchers with a stronger framework of identifying a) what the parameters and components of this trend are, b) who is interested in it, c) who leads opinion, d) how relevant issue-level information is shared, e) which issues shape this idelogy. Moreover, the Twitter-led (or alternative digital trace data, perhaps 4Chan and Reddit) public opinion research approach will have been necessary in the first place to identify the existence of this particular phenomenon. In contrast, a standard-issue public opinion survey aimed at eliciting voting intention would have a hard time illuminating these constituent factors of what constitutes the alt-right, while a poll aimed at social attitudes may indicate some 'comorbidities' of sympathising with the alt-right, but without prior knowledge of the phenomenon, would be unable to pin-point their causes. Besides - in the specific example of the alt-right, a very much online-first ideology, the likelihood of capturing its sympathisers with existing sampling techniques without prior knowledge of the existence of this group would be significantly harder than by using Twitter, or other social media data.

I suggest that non-probability samples with low sample sizes are a serious problem for survey research going forward. What if, out of a sample of 100 African-American voters in a sample of 700 respondents total, 15 lie about having voted in the past election and a further 10 lie about their electoral preferences (social desirability bias), while a further 7 mis-remember whom they voted for in the last Congressional election, and it also so happens that those 7 respondents make up 50% of all sampled African-Americans over 50? While a better (read: larger) sample would have likely helped alleviate such problems, the issue remains that a large amount of information about an important sub-group is at best incomplete. I argue that such scenarios are becoming more prevalent when conducting surveys in the age of declining response rates, and even if sophisticated post-stratification is able to alleviate such problems, it does not address the issue of respondents lying and mis-remembering in non-random ways, and constructing *ad-hoc* opinions based on little beyond interviewer / questionnaire stimuli.

On the other hand, public opinion research derived from digital trace data would be faced with a similar and yet orthogonal problem: assuming there are sufficient amounts of active, tweeting African-Americans over 50 on Twitter, how is one to 'find' them in the endless sea of tweets, without knowledge of individual-level demographics? Such characteristics can be computationally predicted, but it would be disingenuous to claim that this is any 'better' than a flawed sample. Moreover, it is likely that the over-50s recruited for a survey are more like the typical over-50 person than one discovered in a Twitter collection.

[7] with the noteable exception of election results, more on this later

Furthermore, it is important to take into consideration the sustained decline in survey response rates. In practice, this means that it is becoming increasingly more costly and difficult to obtain a random sample of respondents, but it also means that there is likely a biasing factor in *who responds* - they are the same people that sign up for online panels, or go to political events: in other words, a systematically skewed subset of the population. Assuming this decline continues further in the future, statistical methods of imputation and de-biasing using other data, such as government records - or more likely, historical survey data, potentially aggregated - will become more complex, if not impossible. Perhaps this offers an opportunity for linking social media-based approaches with online panel approaches, as both are non-random non-probability samples, but complement each other nicely. For instance, Twitter allows for a much more nuanced and detailed understanding of what the current 'conversation' is actually about; which topics are trending. On the other hand, linking this conversation to stance/sentiment, or any kind of evaluative judgment at the individual level is an exceedingly difficult task when relying only on Twitter data: issues of scale (not every individual tweet can be read, let alone labelled/categorised), comprehension, and missing/unobserved data (what about all the people who do not tweet) are difficult to overcome, so treating Twitter as an arena akin to a focus group (see e.g. Lin et al., 2013) and then feeding those insights back and forth between a survey environment and the social media environment may be a strategy worth investigating in future research. Previous research has linked social media and survey data, such as the British Election Study or several studies at NYU's SMaPP Lab, allowing for both the expansion of data studied, and the validation of measurements from both arenas. It will be interesting to see how surveys can benefit from Twitter-based research, and vice versa, in the future

Survey research has produced a wide array of important research findings across disciplines throughout the 20th and 21st centuries. Furthermore, there are countless experts who, working in academia, the survey industry and government, understand the intricacies of survey design and are able to construct high-quality research designs, even under considerable resource constraints and the other considerable issues facing this paradigm in 2021. But it is nonetheless also true that, especially in the realm of public opinion measurement, a recent overall trend indicates a declining ability of survey data to provide reliable and valid data in the sense that it accurately measures an unambiguously true outcome quantity, i.e. eventual election results. Furthermore, the fact that contemporary survey research no longer necessarily conforms to its own previously held gold standard - random probability samples of target populations - in favour of machine learning, simulation or data imputation, signals a decrease in transparency regarding methodological approaches. Regardless of expected data quality, the cost of conducting a scientifically useful survey of large-scale populations is now so prohibitively expensive, that e.g. graduate students and junior faculty are effectively barred from conducting survey-based projects without the support of entire university departments or private sector partners. Given these trends, there is a strong argument for considering alternative and complementary approaches to survey data for social science research, namely ***observational data*** created by their subjects.

There are multiple types of observational data to draw from. Government and other official records make for very useful data sources, even at the aggregate level. Examples for such data could be: local rent/house prices, homelessness statistics, police report data, tax revenue data, census data (this is essentially survey data, minus the sampling issue), economic indicators, sensor-derived data, or vehicle registration data. Many such data can be used as proxies for social phenomena of interest. Government data, however, are not usually accessible to researchers without extensive negotiation and good will, and further, researchers may question the trustworthiness of governments in many countries, and thus the reliability of data they generate. However, assuming that governments, especially those at regional and local levels have their citizens' best interest at heart when generating actionable data, they should in principle generate valid and reliable insights.

More importantly however, both for this thesis and the current and emerging social science research landscape, I argue that *digital trace data* should form a key alternative and supplementary role in social science research. There are multiple kinds of digital trace data: WiFi hotspot check-ins, social media posts, generic social web interactions (e.g. likes, comments, social payments, group activity), credit card transactions, browsing/cookie data, GPS-enabled app data or social product and site reviews, to name but a few. What unites these types of data is that they were generated - not necessarily knowingly - by an individual whom they refer to. Hence, they are observational data, in that they document a given behaviour by a given individual at point t. In some cases, such as WiFi hotspot check-ins, these data are comparably sparse: they will contain users' device identifiers, a timestamp, and potential other information required to use the service. In other cases, for instance a social media post or interaction, these data are very rich: beyond the content of the post (text, image, video, URL), it will contain information on the users' profile, such as their number of followers and friends, their time zone and language preferences, the number of interactions this post has received, oftentimes specific entities referenced in a post (such as hashtags or user-mentions), and even machine learning-derived descriptions of the content contained in a post (Dellinger, 2018). This broad portfolio of content- and user-specific metadata allows for a deep and insightful study of social phenomena. However, with many kinds of digital trace data, researchers do *not* have immediate access to such data, and requesting access may prove costly or impossible. However, it is important to note that these data do *exist* and hold a great potential for illuminating individual-level behaviour in the contemporary world. It is important to note that, besides the problem of availability, digital trace data come with their own set of problems, and are certainly not free of bias. For instance, is a tweet outlining an individual's views on President Trump's tax cut bill free of social desirability bias, or rather *more* prone to it than an anonymous online survey, given that it will be read - and judged - by a potentially huge audience of peers and adversaries on the internet? There is evidence suggesting that this is indeed the case (Das and Kramer, 2013), and it is also intuitively plausible, given the incentive structures surrounding one's proliferation on social media sites, and especially Twitter.

So, while a tweet offers a datum created without any outside intervention, and thus may on its face seem more intuitively trustworthy and valid than a survey response generated as a result of a stimulus, the existence of the tweet tells us nothing about the motivation and the circumstances influencing its generation. What if the author's peer group implicitly influenced them to publish this tweet (and thus create the datum), even if they do not necessarily believe in what they posted? Moreover, the questions of sampling and representativeness, and indeed the categorisation of individuals as members of (demographic) groups, or indeed as humans when using digital trace data for public opinion research. This highlights that this paradigm is by no means less prone to biases, shortcomings or trade-offs when compared to survey research. Indeed, it may even be *more* prone to those that exist for surveys, and also far from ideal in other regards, such as the governance of data access or platform rules.

Nonetheless, I have outlined the increasing difficulty of conducting high-quality survey research, and unlike with surveys, which have been the subject of large-scale research and improvement for decades, I argue that the digital trace data landscape, more specifically the field of Twitter-based public opinion research, offers countless avenues and ample scope for exploration, basic research and improvements. So, ultimately, the goal of such improvements should not be to replace surveys - surveys will continue to be hugely important tools in the arsenal of social researchers across the globe, and a gold standard which other, emerging approaches will have to be measured against. Instead, a maturing field of public opinion research with digital trace data can help address the gaps that surveys may be less adept at filling in the 21st century.

In this section, I have highlighted the core problems that exist with the survey research paradigm that has led social inquiry for the past century. While I do not claim that survey data is inherently unreliable and invalid, there exist significant issues regarding the artificial setting of the data collection process, the treatment of answers as 'truth' and the real-world constraints facing survey practitioners. Further, I have introduced alternative types of data that could be used for survey research in general and public opinion research more specifically. I revisit the most promising of such alternatives, digital trace data in the form of tweets, in the final section of this chapter.

2.3 Forecasting and prediction

Besides public opinion, this thesis refers frequently to the concepts of ***prediction*** and ***forecasting***. Both are crucial to the process of scientific research. Whenever research establishes a relationship between certain factors, or variables, we may want to know how we can expect these variables to affect certain outcomes given certain conditions. Similarly, science is often concerned with forecasting: If it is known that a given event E will take place in the future, and we also know certain conditions which may have been predictive of E in the past, we can now forecast the nature of E at time t given this knowledge. While everyday language typically does not draw a clear distinction between the two terms, subtle differences do exist. Generally speaking, prediction refers to the estimation of relevant quantities given other factors/variables. These quantities may be hypothetical, past, or future. Predictions can be qualitative or quantitative in nature. Understood from a scientific perspective, prediction can often be used interchangeably with the term 'modelling' - variables are observed, and based on the nature of their (co-)occurrence, are used to 'model' a variable Y. Forecasting, on the other hand, while a sub-type of prediction, is concerned with *future* events, and considers the temporal nature of events as necessary for forecasting. Hence, weather forecasts rely on historical weather data as well as present weather data to inform weather forecasts at time t. A weather prediction model may only use other indicators, such as ambient pressure, precipitation, etc. to predict, and thereby *understand* and explain weather patterns. However, it is important to note that these definitions are not canonical, and are, even in the academic literature specifically concerned with studying its intricacies, frequently conflated and used interchangeably (see e.g. Rescher, 1998; Armstrong, 2001).

This thesis employs both concepts defined above. Specifically, I employ the forecasting and prediction of *election results* as a way of further investigating the feasibility of measuring public opinion as it is expressed by individuals by posting to Twitter. This means that this research seeks to understand the possibility and logistics of predicting the outcome of an event *E* (in this case, a given election) taking place at time *t*, at a given time point *t-n*. At this stage it is important to reiterate that this thesis focuses on forecasting election results for instrumental reasons rather than in their own right, as election results offer unambiguously true outcome variables of events shaped by public opinion, the substantive area of inquiry this thesis contributes to. Hence, the advances to forecasting election results from Twitter data outlined in this thesis should be understood as primarily contributing to the study of public opinion, rather than to the election forecasting literature per se.

2.3.1 The art and science of forecasting and prediction

> *"Everything we do involves forecasts about how the future will unfold. [...] The problem is, we're not very good at it"* (Tetlock and Gardner, 2015, p. x2)

Predicting future events in order to inform behaviour and decision-making is an essential component of the human condition - a process as potentially mundane as setting out which clothing to wear based on a brief look out of one's window, to life-and-death questions of high politics. In other words - "human life is 'futurition', largely determined by what is not yet realized" (Rescher, 1998, p.2).

While it may have become commonplace for people to be able to watch televised weather forecasts and structure their days accordingly, or read about election polls that come within decimal-point ranges of election outcomes, forecasting approaches in varying fields have progressed from pseudo-scientific pursuits to robust, scientific approaches reliant on replicable algorithms and mathematical/statistical methodologies underlying the principles of probability. However, both the examples of weather and polls are frequently inept at producing consistently accurate forecasts: polls often fall short of accurately predicting election outcomes, while weather forecasts, especially when covering a time-period longer than a few days, are notoriously unreliable. In short, forecasting the future is by no means *easy*, regardless of which event space the forecast operates within.

When considering the transition from pseudo-scientific to systematic forecasting, meteorology functions as an excellent primer, as "meteorology leads the forecasting world" (Lewis-Beck and Stegmaier, 2014, p. 322). In their 2014 essay, electoral forecasting pioneer Michael Lewis-Beck and his colleague Mary Stegmaier provide a brief overview on the development of this discipline, and the potential lessons other prediction-focused fields may be able to draw from it. The authors identify the Norwegian physicist Vilhelm Bjerknes's 1904 proposal for increased reliance on large-scale *observational data* and a greater understanding of the atmosphere - in other words, the system wherein the event which is to be forecast takes place - as a key trigger for the introduction of the era of scientific weather forecasts[8]. Without sufficient knowledge of the atmosphere - or, to generalise, *the system encompassing the event's constituent components* - and ample observational data, forecasts will fall short of the target - congruence with later events. Furthermore, as Rescher (1998) states - "the relative stability of the relevant factors is thus crucial for prediction. And this means that local, problem-specific circumstances will be determinative" (p. 79), meaning that events in a system with large-scale volatility will be exceedingly hard to forecast.

[8] Notwithstanding the fact that Vilhelm Bjerknes's equations turned out to be incorrect once tested

The understanding that successful forecasts rely on large amounts of *observational* data supports the case for an investigation of the feasibility of systematic election forecasting using Twitter data. The fact that they are plentiful and observational helps the case, but, as discussed throughout this thesis, knowledge of the system within which they operate, and especially the relevant sub-system which contains signals pertaining to public opinion considerably lags behind the availability of data.

It is further interesting to consider the role of prediction and forecasting in science more generally. While they are typically seen as necessary in the ’hard’/natural sciences, opinion is divided regarding the ’forecast-ability’ of events and phenomena studied by the social sciences, leading many to consider social science as not truly scientific and even imprecise. While there is clearly ample evidence to support the bad track record of social scientific forecasts, there is a recent development which very much brings under scrutiny the idea that social science is inherently unable to forecast and predict, namely the emergence of *big data*. I understand this term as synonymous with the emergence of *digital trace data*. Human online (and in some cases offline) behaviour is now documented and often uniquely identifiable to its source, the individual. And while it may not be considered ’social science’, as it does not result in papers based on testable theories and hypotheses, the largest corporations in the world are engaging in constant, highly sophisticated forecasting and prediction of human behaviour: Amazon recommends items based on customers’ previous purchases and traces they leave while browsing the web. Facebook prioritises content they think will keep users on the site, while simultaneously bombarding them with relevant, user-specific advertising. YouTube auto-plays videos their predictive models identify as being most likely to keep users on the platform, with a promise of ever-better recommendations, if only the user keeps consuming more content (and advertising). Cities and municipalities on the other hand are using digital trace data generated by their citizens - be it through individual-level interaction with online services, or data collected through sensors - to improve their services. Examples for this can be subsumed under the ’smart city’ moniker, and could include dynamic, adaptive traffic signalling systems, adaptive, ’smart’ electricity grids, or demand prediction models for shared resources, such as smart bike sharing programs, whereby forecasts are made regarding demand at a given location at a given time, and resources are allocated accordingly. If construed in a more sinister fashion, one may argue that cities and other public sector institutions use big data as a means for surveillance and control - current examples of this include facial recognition technology paired with individual-level location and movement profiles, or ’predictive policing’, which uses digital trace data to pre-emptively allocate resources to certain areas and groups considered as most susceptible to criminal activity. Beyond this, electoral campaigns make use of big data by targeting their online advertising to the precise individuals they identify as essential for victory.

All of these examples - corporations, political operatives and public entities - have realised that the emergence of big data, coupled with continuing advances and developments in machine learning methods and computing capacity, potentially allows for more accurate predictive social science. This produces strong incentives for further investment in these areas, as more forecasting and prediction yields yet more data, and thus implies iterative improvement of future models, and increased utility (read: profit) of any given data point. The utility of big data for forecasting has not, however, necessarily made its way into the mainstream of academic social science. While there has been a significant uptick in published research that avails itself of digital trace data, forecasting has been limited. Overall, however, I argue that the lack of adaptation of digital trace data in empirical-predictive social science stems from the simple fact that these data are, on the whole, *not available* to researchers. So, data scientists choose *not* to work at universities and research institutes, but instead go to work for large corporations, where they *do* have access to these data, and are able to build comprehensive forecasting models, the likes of which social scientists of yesteryear could only dream of. In essence, then, I argue that the conditions for successful forecasting in social science exist, but academic social scientists may not be the ones mainly conducting them at this moment in time.

2.3.2 Forecasting election results

Conventional wisdom may suggest that 'electoral forecasts' are synonymous with sample survey polls. Undoubtedly, they are the most common means by which the general public's mood toward available electoral choices is measured and reported, be it leading up to elections, or simply to gauge the current political climate. However, it is somewhat misleading to describe opinion polls as *electoral forecasts*. Unless questions are phrased in an unusually distinctive fashion relating to a respondent's knowledge rather than her opinion - e.g. 'Who will win the election in November?' or 'Which percentage will candidate C receive in constituency D?', opinion polls capture the aggregate mood of the electorate - or any population being surveyed - at a given time t. Hence, what polls really do is *nowcast* rather than *forecast* - when the poll is closer to the date of the election, the substantial meaning of the two terms converges.

The following section outlines different approaches to understanding how elections will turn out before the fact - the field of election forecasting. The initial focus lies on the history and rise to power of scientific sample-survey polling, the first systematic means by which election results were able to be forecast. Furthermore, I introduce different approaches to producing systematic predictions of election outcomes developed and discussed in the political science literature, in particular relating the approaches to the framework for successful forecasts introduced above.

Polls: quantifying the public's mood

How did elections even 'work' before polls? To trace back the origins of electoral polling, the widely read magazine 'Literary Digest' is a good starting point. Even though some scholars see the quasi-polls conducted in North Carolina and Delaware in 1824 (Smith, 1990) as the original 'proto-polls', The *Digest* and its reader straw-polls are widely considered one of the first large-scale attempts at forecasting election outcomes by means of surveying a sample of respondents[9]. They got it right on four consecutive occasions - Harding, Coolidge, Hoover and Franklin D. Roosevelt were all proclaimed by the magazine as likely to win the U.S. Presidential elections, based on their large-N sample of what the *Digest* referred to as 'voters', most would however consider 'readers'. Retrospectively, some might suggest that this simplistic, bias-prone methodology was destined to fail - and fail it did, at predicting FDR's re-election in 1936: despite the fact that their sample was said to comprise around 2

[9] Several newspapers and magazines had previously employed straw polls to make electoral predictions, however, their readership restricted potential sample sizes and implied that forecasts typically referred to a given, small-scale geographic entity, e.g. a district or a county

million respondents, it did not accurately represent the target population. Robinson (1932) produced one of the first academic appraisals of the efficacy of sample-based straw polls in electoral forecasting[10], arguing that the *Digest* fared well in its first four attempts due to the wide margin between actual votes received by candidates. In other words, had these elections been 'too close to call', the *Digest's* poll's inaccuracy at every level below that of the national-level popular vote would have exposed their flawed methodology earlier (p. 59) - as turned out to be the case in 1936 (Crossley, 1937, p. 26). Put simply, *Literary Digest* and other straw poll 'pioneers' failed to account for bias in their samples. By only sampling from their readership, they failed to account for variability in voter turnout and also failed to account for demographic and socio-economic variability throughout the United States (Crossley, 1937).

However, the 1936 U.S. Presidential election proved important to the electoral forecasting space besides the *Digest*. It also saw the emergence of George Gallup and contemporaries Elmo Roper and Archibald Crossley. Gallup introduced a method he referred to as 'quota-control' (Crossley, 1937, p. 29), something most pollsters and statisticians would today refer to as *weighting* and a form of *stratified sampling*. In other words, Gallup ensured his sample was stratified by known socio-demographics. By doing this, his vote-distribution predictions for the electoral college were far more accurate than that of *Literary Digest*, and he got the eventual winner right.

Gallup's 1936 team updated their projections as new data came in. Hence, this is an adequate example for explaining the difference between what is today sometimes referred to as *nowcasting* as opposed to the concept of *forecasting* - Gallup's first poll was likely less predictive than the final one. In other words, polls capture the *current* mood of a target population, as if the election were to take place there and then, rather than its mood in $t + n$[11]. In the modern-day political landscape, full of 24/7 news coverage and new polls on a daily basis, poll numbers can shift significantly in a short time period, meaning final election results can diverge considerably from polls published at t_0, making for the following formal expression of the difference between a nowcast (N_i) and a forecast (F_i, either poll-based or otherwise) of an election E_i taking place at time t_1 when regarding the contemporary polling landscape, assuming all instruments are both valid and reliable in measuring what they aim to measure:

$$N_i(t_0) \neq E_i$$

$$F_i(t_0) \approx E_i$$

$$N_i(t_1 - x) \approx F_i(t_1 - x) \approx E_i$$

Furthermore, evidence suggests that the abundance of polls in the modern era, twinned with the sometimes sensationalist way in which media outlets report on them, may indeed be contributing to higher levels of voter preference volatility in the modern day (see e.g.

[10] As well as other contemporary forms of forecasting at the time, which one would today likely refer to simply as "punditry", with a success record to match this discipline (see Robinson, 1932, p. 1-45)

[11] Nonetheless, one can confidently assume that the observed divergence between a nowcast conducted at time t_0 (e.g. four weeks before the election) and a nowcast conducted at time $t_1 - n$ (e.g. two days before the election, so $n = 2$) would have been lower in Gallup's day, barring any major political scandals or upsets outside of the norm of the news coverage of electoral campaigns than it would be today, thus allowing the classification of Gallup's t_0 polls as quasi-forecasts.

McAllister and Studlar, 1991; Rattinger and Wiegand, 2014). To sum up, it is important to take note of the distinction between polls as nowcasts (with ample time before the event takes place) and polls as forecasts (very shortly before the event).

Since these formative days of election polling, it has undergone key technical innovations, such as the widespread introduction of computer-assisted telephone interviewing (CATI) technology or internet surveys. Yet, at its core, the key to polling's success still rests upon two key factors: the representativeness of the sample and the degree to which pollsters accurately estimate turnout. Hillygus (2011) identifies pollsters' shortcomings in modelling turnout as a key factor in the "polling debacle" (p. 965) of 1948, where Gallup forecast Dewey to beat Truman to the U.S. presidency by a margin of 5 percentage points. This is where Gallup's quota sampling method reached its limits: targeted sampling, rather than post-stratifying a random sample leaves inclusion and exclusion criteria at the discretion of the pollster - and, in this case, the forecast's poor showing suggests that Gallup's information on how the sample should be composed was flawed. This polling debacle of 1948 ushered in a new gold standard in polling, the random sample (see e.g. Mosteller, 1948).

To sum up, polling is still the most widespread tool in the space of electoral forecasting, despite most published polls not being forecasts in the true sense of the term.

Political Economy & statistical models

Forecasting approaches employing variables other than voters' self-disclosed voting intentions in predictive statistical models, such as macro-economic indicators or subjective economic indicators offer an alternative to polling, which is firmly grounded in the political science literature and functions rather differently than polling, in that they can be conducted with ample lead time, no need for polling data, and a parsimonious set of explanatory variables.

The most prominent type of electoral forecasting models employ economic variables; the broad academic consensus being that economic factors, both objective/macro-level (such as GDP growth, unemployment or inflation rates) and subjective (one's personal perceptions of the state of the economy) have a large-scale, significant impact upon voters' electoral choices (see e.g. Fair, 1978; Tufte, 1980; Lewis-Beck, 1988; Fair, 2011), thus mirroring the much-quoted phrase 'It's the economy, stupid!' originating from Bill Clinton's 1992 presidential campaign. The key underlying idea of such economy-based models argues that voters either reward or punish the incumbent party's candidate for their management of the economy, whereby punishment results in an increased number of votes for alternative candidates or parties (see e.g. Graefe, 2013, p. 2). This approach is based on the assumption that voting preferences are primarily shaped *retrospectively*, i.e. by means of evaluating the incumbent's performance, rather than *prospectively*, whereby future developments and campaign promises are taken into account. Indeed, Miller and Wattenberg (1985) found that "incumbents have been judged primarily on the basis of retrospective performance, challengers on prospective policy, and candidates running in non-incumbent races on prospective performance" (p. 359), thus suggesting that economy-based forecasting models may be oversimplifying the complex phenomenon of voters' electoral preferences. However, the track record of such models is overwhelmingly positive when judged by their ability to accurately forecast the *winner* of an election. In the case of U.S. presidential elections, 4 out of 40 published ex ante political economy forecasts between 1996 and 2012 (5 elections) failed to accurately forecast the respective winner. The mean absolute error for all forecasts was 3.1 % (Graefe, 2013, p. 2-4) when referring to vote share forecasts.

An early example of an applied political economy vote-forecasting model is featured Ray Fair's (1978) "The Effect of Economic Events on Votes for President". Using four macroeconomic indicators, annual GNP per-capita growth rates, annual GNP deflator growth rates, annual unemployment rates and quarterly election-year GNP growth rates, he estimates the parties' vote share using several regression models, whereby GNP growth rate proved to be the best fit for explaining vote share (p. 167). Building on this work, Fair developed the following equation, whereby G is the GNP growth rate, P the GNP deflator, V the incumbent party vote share, Z the number of quarters in which the growth rate exceeds 3.2%, $DPER$, a party dummy, DUR, the length of incumbency, and WAR, a dummy which equals 0 if the election takes place after 1948:

$$\begin{gathered} V = 47.75 + 0.667(G * I) - 0.690(P * I) + 0.968(Z * I) \\ +3.01(DPER) - 3.80(DUR) - 1.56(I) + 4.89(WAR) \end{gathered} \tag{2.1}$$

In his original paper, Fair's model predicted vote share with an error term ranging from -.007 to -.071 (p. 168) for a total number of 16 elections. This early application of an economic model of electoral forecasting provides evidence for a strong linkage between voters' electoral choices and the objective economic landscape, despite unexplained factors that can be attributable to the variability in error size. Furthermore, Fair improved on the performance of his model in later iterations (see e.g. Fair, 2011).

Michael Lewis-Beck has been a further proponent of forecasting elections well ahead of time using economic data. Unlike Fair, Lewis-Beck emphasises the importance of an indicator of a political nature in his models, typically a certain incumbent popularity measure. From this follows his general equation:

$$Vote = f(politics, economy) \tag{2.2}$$

His key model (Lewis-Beck and Rice, 1992) argues that (future) incumbent party vote share in the U.S. presidential election can be modelled as a function of the sitting president's approval rating (PP) at a given time t, the percentage change in GDP from the fourth quarter of a given pre-election year to the second quarter of an election year (G), the number of House seats the incumbent's party gained at the last election (PS) and the support for the presidential party in primaries (C). Furthermore, Lewis-Beck developed comparable models for different types of U.S. elections and elections in other countries. Crucially, Lewis-Beck's forecasting models, as well as his other work on economic determinants of voting behaviour rely on the 'Responsibility Hypothesis' (Lewis-Beck and Stegmaier, 2000, p. 114), whereby voters are understood to blame or reward the government for the trajectory of the economy. Lewis-Beck and colleagues have applied this and similar methodologies in several publications (see e.g. Lewis-Beck and Rice, 1984, 1992; Lewis-Beck, 1986, 1990; Bélanger et al., 2005), with a strong degree of forecasting accuracy, when judging the models on their effectiveness in correctly forecasting the given election's winner.

Other scholars have argued for the primacy of political factors (Powell and Whitten, 1993, see e.g.) or individual-level factors of 'political sophistication' or sociotropic preferences (Gomez and Wilson, 2001) as more salient shapers of voting intention than economic factors and hence the optimal variables in predictive models forecasting election outcomes. Graefe (2013) developed a forecasting model focused on campaign events, candidates' perceived leadership qualities and measures of party identification, thereby outperforming or matching the accuracy of all political economy models between 1996 and 2012, as well as beating the final Gallup poll when using pre-election day data.

Overall, there is evidence in support of both economic voting models and non-economic models. Anderson (2000) suggests that both institutional and political factors mitigate economic effects, and Erikson and Wlezien (2012), widely considered as authorities on the study of the determinants of voting choice, argue that "[t]he exact voter psychology by which voters respond to the economy remains unclear" (p. 110). In other words, a multitude of approaches may be valid in a given context to which it may be applicable, and, in any case, this literature review is not set out to evaluate the merits of every forecasting model's underlying theoretical assumptions. Nonetheless, it is important to note that, besides proving highly predictive of eventual election results, the statistical election forecasting literature, especially early work by Tufte and Fair provided evidence on the key macro parameters determining the outcomes of elections, thereby significantly influencing other sub-fields of political science and political economy.

I argue that this widespread contribution to the field beyond 'How to forecast elections: a guide' forms the most important takeaway from the statistical election forecasting literature for this thesis. By instrumentally using the scenario of forecasting elections, with data not necessarily immediately linked to the desired outcome variable, the proponents of statistical election forecasting were able to learn more about a different, yet related subject, namely the determinants of electoral outcomes, in general. In essence, this is an endeavour I am seeking to replicate in this thesis, by transposing from macro-economic indicators to Twitter data, and from learning about elections to learning about digital trace data and public opinion. Hence, I am not necessarily borrowing methodologically from this discipline (although, of course, these findings inform the research of anyone working on elections), but rather taking inspiration from their rationale and the success of their pursuit. Hence, is is important to reiterate that, I am less concerned with matching polls or statistical models' ability of correctly forecasting election results. Rather, I am interested in deepening the understanding of how digital trace data can be used in social science research. Election results provide a rare occasion in the social sciences where there exists an unambiguously *true* outcome variable to which predictive/forecasting models can be benchmarked and iteratively improved. In other words, my forecasting attempts are instrumental to furthering the understanding of how digital trace data can be employed in social science research, and crucially, how public opinion can be extracted from them.

2.4 Digital trace data in academic research

In the final section of this literature review, I pivot away from traditional (offline) work relating to election forecasting and public opinion, and focus instead on recent advances in social science research availing itself of user- and event-level digital trace data. Such data are digital records of humans' interactions with contemporary connected technologies, be it social media, cashless payments, individuals' traces when browsing the web or data collected when humans' mobile devices connect to the internet, such as sensors or public

WiFis, to name but a few. Furthermore, it is important to reiterate that the majority of such digital trace data are completely proprietary, and will neither be accessible by the individual who created them, nor by researchers intending to learn more about the world, nor by governments seeking to regulate services or prevent/investigate criminality. However, large amounts of research is being conducted in order to learn more about human behaviour in the digital era - most any company that deals in ’data’ (and this is, in 2021, most any company on the internet) will have a team of data scientists and analysts who mine the data generated using their products and services to understand more about their user-base, and to increase efficiency and thus profitability. In the case of the largest companies of the 21st century - Facebook, Amazon, Google, etc. - this amounts to impactful evidence on general human (technology-mediated) behaviour. However, this evidence will only be shared with the world at large if it supports the company’s bottom line. Other research questions, which would require such companies to share their data are either not being investigated, or the evidence resulting from them is not being shared with the public, as companies may understand their impact to be harmful to their interests. Such questions may include: How does YouTube’s recommendation algorithm work? How are individuals radicalised and misinformed through WhatsApp? How does engagement-driven content delivery on Facebook, Instagram, etc. affect human cognition, mental health, and other psycho-social attributes?

However, this is not the case across the board. Some owners of large-scale digital trace data have realised that the benefits of sharing outweigh the costs. This includes, for instance, public transport systems around the world, such as London’s TFL (tfl.gov.uk, 2020) or New York’s MTA (mta.info, 2019), whose decision to make large amounts of dynamic and historical data available to developers and researchers through an API has allowed for the development of several critically acclaimed public transport applications, such as CityMapper, which allows users to optimise their commutes and reduce their carbon footprint. The same is true for many other (inter)governmental agencies/organisations around the world, who provide large-scale access to aggregate data on the web, such as data from the US census bureau or the World Bank. However, crucially, in both cases, these are *not* individual-level data generated by the user accessing them, but rather aggregate data collected and generated by the government. When it comes to individual-level data, individuals are heavily restricted in what they can access. Indeed, the least restrictive platform here is Twitter. It allows any user who registers for a developer account access to (potentially) *all* of Twitter, by means of retrospective search (keywords, hashtags, etc.) and dynamic streaming collections. The stream is limited to a maximum of 1% of the entirety of content published on Twitter, whereas this is not applicable when keywords are so narrowly defined that they do not cover 1% of the entirety of Twitter at any given moment. For this reason, Twitter has become the de facto default data source for researchers interested in working with individual-level digital trace data.

Crucially, Twitter data, and any digital trace data captures records of human behaviour *only* when it is mediated by technology. If the social science community understands this as simply equating to ’human behaviour’, we are likely making a grave error, as a) much of human behaviour still occurs wholly unmediated by technology, or is at least not captured by Twitter and b) even for much of the behaviour that is technology-mediated, not all aspects of any given behaviour are recorded as digital trace data, or indeed even recordable. Hence researchers must always keep in mind the potential of bias, and keep in mind that any digital trace datum only exists because someone designed it to be recorded and re-recorded in precisely that way. It is further important to keep in mind that the generation, let alone the dissemination, of digital trace data is entirely at the whim of the companies that own them, and likely downstream of their profitability. I argue that knowledge of this is essential when using digital trace data in social science research. In practice, this results in a number of useful considerations when employing digital trace data (and especially social media data) in social science research: 1) human-computer interaction is influenced by the system design of the platform on which it takes place, and the data produced by it will be biased in this way - the instance of human behaviour that created the datum would not have taken place were it not for the platform, and if the platform were structured even minimally differently, the datum would be different, even if the behaviour may have been essentially the same. 2) a 'like', 'follow' or 'retweet' means nothing on its face, but rather requires interpretation on a case-by-case basis in order to ensure whether it actually signifies the user's approval of a given subject/topic of discussion . 3) the vast majority of contemporary data governance is *not* democratic, and given the understanding that digital trace data are the oil fuelling the information age, researchers should not underestimate the skewed incentives that exist at every stage of the data production, dissemination and analysis process, and, in my opinion, should argue strongly for democratic data governance coupled with enhanced individual civil liberties in the information age.

Nonetheless, I now introduce existing research availing itself of such types of data, as the imperfect should not be the inhibitor of progress, and, alas, this thesis is situated firmly in the camp of using digital trace data - in this case tweets - as they are, not as we may wish them to be. Initially, I focus on general digital trace data literature, after which I discuss the Twitter-based public opinion and election forecasting literature. It is important to note that this section of this thesis' literature review is by no means extensive, but rather functions as providing the reader with an overview. The individual literature reviews in this thesis' three self-contained papers provide additional depth and detail on the specific areas of previous research employing digital trace data which pertain to these papers' specific subject areas.

2.4.1 (Social science) research using Twitter and other digital trace data

There is a growing body of research which has used Twitter, other social media data and digital trace data more broadly. In Public Health, tweets have been used to predict the spread of swine flu (Ritterman et al., 2009), further understanding of disease concern in the public (Barros et al., 2018), characterizing "diabetes, diet, exercise, and obesity" (Karami et al., 2018), or developing dynamic predictive models for asthma-related emergency room visits in the USA (Ram et al., 2015). Furthermore, researchers have also employed "Wikipedia access logs" as a means for disease monitoring (Priedhorsky et al., 2017). In criminology, scholars have used Twitter data to build models dynamically predicting the occurrence of violent behaviour (e.g. Wang et al., 2012), while others have sought to dynamically track the incidence of illicit drug use (Buntain and Golbeck, 2015; Buntain et al., 2015). In economics, tweets have been used to forecast the trajectory of stock markets (Bollen et al., 2011a; Si et al., 2013; Rao and Srivastava, 2012; Zhang et al., 2011) with differing levels of success. In geology, Twitter data have been used to develop dynamic earthquake tracking models (Sakaki et al., 2010), while in geography, Twitter data have been used to map the geo-spatial incidence of and spread of different languages around the globe (Mocanu et al., 2013).

In political science, the use of social media data has been particularly prevalent; be it investigating twitter users' ideology (Barberá, 2015; Bond and Messing, 2015), the effect that exposure to political tweets has on individuals' political knowledge (Munger et al., 2016), the phenomenon of group organisation for political protests using social media (Theocharis et al., 2015), or the determinants of agenda-setting and information flows in US politics (Barberá et al., 2019) . More recently, there has been increased attention on the spread and impact of 'fake news' / mis-disinformation through social media and its impacts on (electoral) politics (Allcott and Gentzkow, 2017; Guess et al., 2018).

It appears that much research availing itself of digital trace data has several traits in common, regardless of the discipline: researchers tend to emphasise the dynamic nature of (Twitter) data streams, and how this allows for the monitoring and/or prediction/forecasting of a given event/variable/quantity. This is the case for weather or geological events, as well as disease incidence or illicit drug use. While such publications provide logically plausible research designs, I argue that their real-world applicability, especially when used without any other source of outside data, and without large-scale attention to individual-level user traits in order to map signals obtained from platforms to target offline populations, is actually minimal. This is due to a) the vast majority of the world's population *not* being on Twitter, and b) the kind of people who tweet about illness, drug use, or similar things not being representative of the Twitter population as a whole. So, while a Twitter-based monitoring system for earthquakes or similar rare events, as well as ubiquitous and universal events, may be plausible, I argue that it may be less useful for illegal activities, or the sharing of potentially embarrassing information, such as that pertaining to one's health. At the same time, there appear to be fewer published papers that leverage Twitter's giant tweet archive as a means of *understanding and explaining* phenomena rather than dynamically monitoring and predicting them. A great example of one such work is Mocanu et al.'s (2013) paper on the global spread of different languages, a highly creative and suitable project for this type of data. In other words, there is ample room for further exploration of Twitter (and similar) data as vehicles for explanatory (social) science, rather than simply building semi-effective monitoring pipelines.

Overall, I argue that political science has produced creative and useful research in this space, especially when it was concerned with explanation (fake news, individual-level political ideology, protests, agenda-setting) rather than dynamic monitoring. However, this is not to say that dynamic monitoring of anything - in particular, elections and electoral campaigns - is futile. It is more to say that the promise of Twitter's endless data stream as the perfect 'everything-monitoring' solution may be overstated, because of its highly non-representative user-base and the fact that tweeting is entirely voluntary. Furthermore, it is likely somewhat over-simplistic to assume a mere mention of something deemed a relevant keyword by researchers before the fact - 'drugs', 'asthma', 'hospital', 'Donald Trump' - may mean very little in isolation, and cannot simply be aggregated up to build a dynamic model measuring anything beyond the actual incidence of such keywords. However, as I describe in the following section, the Twitter-based election forecasting literature has delved deeper into these questions: What does a mention of m in context C actually mean? When is it positive, when is it negative? How do we know this person's mention actually matters? How do we know this person is actually a person?

In any case, it is important to note that this style of research, be it dynamic or historical, is likely indicative of the future trajectory of a considerably large part of social science research.

With declining survey response rates and an over-supply of qualified researchers coinciding with an under-supply of adequate funding for large-scale research projects, researchers will increasingly find themselves relying on working exclusively with digital trace data. Assuming Twitter keeps its policy of mostly free access alive, this will continue to be the first stop for academics, even though it would not necessarily be their first choice. If this is however not the case, academics will either find themselves begging large platform companies such as Facebook or Google to give them a tiny glimpse of their data, or worse yet, find themselves working in such companies' in-house research centres, without the prospect of freely defining their research agendas or having a non-restrictive pathway to publishing results their employers may not like. Maybe, then, researchers will have to start becoming active in demanding access to individual-level digital trace data, or building alternative, open, inclusive, deliberative and democratic platforms on which humans can solidify an equitable, progressive digital society.

2.4.2 Twitter-based public opinion and election forecasting research

I now introduce the existing literature which has employed Twitter data (and in some rare cases, other social media data) to forecast the outcome of elections. While these publications differ greatly in a multitude of facets, such as the types of elections they sought to forecast or the sampling and data collection strategies that were employed, the crucial distinction lies in the top-level methodology the researchers employed to go from a collection of tweets to an estimate of who will win an election. Hence, I begin by outlining the two dominant methodological strands employed in this space, starting with keyword-frequency volume approaches and followed by sentiment analysis-assisted approaches.

Measuring volume of keyword frequency in tweet collections is arguably the simplest way of measuring public opinion from tweets, and forecasting election results. Whichever keyword (or set of keywords associated with a given candidate or party) occurs most frequently is understood as signalling the candidate/party with the highest vote share. The first (and widely cited) paper employing this approach covered the German Bundestag elections of 2009 (Tumasjan et al., 2010). The authors found that relative mention volume, using party names and party leader names as keywords, predicted the eventual vote shares with a mean absolute error (MAE) of 1.65%, in line with traditional opinion polling. However, Jungherr et al. (2012) highlighted the shortcomings of such as design: If the Pirate Party, which received a vote share of approximately 2% in 2009[12], had been included in Tumasjan et al.'s original analysis, this party would have been the winner and largest party. This highlights one of the key shortcomings of volume-based approaches: *Researchers' design decisions and inclusion criteria have a significant influence on the accuracy of the forecast*. In retrospect however, it is arguable whether a similar design decision would produce similar results if it were repeated today. In the following years and elections, the Pirate Party has become obscure in Germany while Twitter use in the country (and globally) has risen significantly. Perhaps, then, the unique case of the Pirate Party being run and loved by early adopters who were also hugely influential on German-language Twitter at the time, is something that would not repeat itself nowadays, and would not translate to different party systems. At the very least, however, it illustrates a huge complexity of using Twitter as a

[12] and, due to the German constitutional arrangement of a necessary 5%-vote share threshold which needs to be reached in order for a party to enter parliament, did not end up sending any MdBs to the Bundestag

data source - the platform 'evolves' dynamically, beyond not only the control of its managers or individual users, but also beyond their timely comprehension of change. To assume that anything which was 'true' mere years ago can be translated to the current context without considering all available factors influencing temporal decay of research findings drawn from a different state of the platform from which data originate is likely incorrect (see e.g. Munger, 2018)

Other scholars have used similar volume-based designs, however aiming to avoid a similar error as in Tumasjan et al. (2010) by more rigorously specifying keywords, or only tracking hashtags known to be associated with the election or relevant candidates / parties (e.g. DiGrazia et al., 2013; Caldarelli et al., 2014; Cunha et al., 2014; Jungherr, 2013, 2014; Nooralahzadeh et al., 2013), while other authors included volume-only forecasts in their publications in order to compare them with other approaches (e.g. Sang and Bos, 2012).

The most accurate applications of this method achieved mean absolute errors of under 2%, which is in line with traditional pre-election opinion polling (e.g. Tumasjan et al., 2010; Sang and Bos, 2012). However, overall, this approach is unstable and unpredictable in the sense that the conceptual framework of the linkage between observed tweet text and vote choice is a black box (Loynes and Elliot, forthcoming), and accuracy is very much influenced by seemingly arbitrary design decisions taken by the researchers, such as keyword specification, data pre-processing, the timescale of data collection or the sampling frame of collections. As Ceron et al. (2016) put it in their comprehensive book on the state of the field: "no matter whether they record attention, awareness or support, merely computational data seem to retain some problematic attributes and might fail to catch the informational complexity of the social media environment." (p.20). This highlights an interesting question which has been investigated empirically (see e.g. Jungherr, 2014), but not answered comprehensively: what exactly does keyword-derived volume tweet data signify? Most likely, it is a combination of all three - *attention*, *awareness* and *support* toward a candidate or party of interest. However, besides support it may also contain its inverse: dislike or disapproval as directed towards a candidate or party, which is an idea leveraged when employing sentiment analysis to better measure support within tweet volume.

To sum up, volume-based methods offer a tantalisingly simple means of producing Twitter-based election forecasts, but: they do not consistently work. However, volume data *may* have a useful application in Twitter-based election forecasts, and Twitter-based public opinion research in general, when they are understood as a measure of attention/awareness (pure volume) or support/dislike (positive/negative sentiment volume). This way, they could be used to help adjust and weight individual-level data. Furthermore, tweet volume is a useful tool for tracking campaign-relevant events in quasi-real time ('nowcasting'), thus potentially highlighting shifts in trends, which can again feed into more sophisticated and theorised research designs.

However, the problem of tweet-volume forecasting goes beyond the lacking understanding of what keyword volume actually signifies in a large-scale, aggregated, *political* context. Ultimately, the hugely un-representative nature of Twitter as a platform - it is home to disproportionate amounts of highly educated, urban, young, white and male users (Rainie, 2012) - makes it difficult to derive any meaningful insights from tweet keyword/hashtag volume, even assuming we knew exactly what it signifies. In other words, we may know that people on Twitter like/dislike/are aware of candidate C or event E, but how can this be mapped out to a larger target population, some of whose elements are at best under-represented and at worst not represented on the platform at all? Gayo-Avello (2013) highlights these shortcomings, and calls for more basic research into the socio-demographic composition of the platform, and the ways in which it differs from the general population. I have taken this research agenda to heart in this thesis, and discuss my targeted contributions in this problem space in the last section of this chapter.

Besides volume-based approaches, the other highly prevalent methodological approach to Twitter-based public opinion research in the form of election forecasting has been ***sentiment-based forecasting***. Sentiment analysis is a methodology of computational linguistics which seeks to measure the 'sentiment', i.e. the emotional direction of a unit of natural language text using computational approaches (see 'Understanding Political Sentiment' for an in-depth review of the literature on sentiment analysis). If large-N collections of tweets concerning political candidates in elections are indeed indicators of "attention, awareness or support" (Ceron et al., 2016, p. 20), then the goal of incorporating sentiment analysis into election forecasts can be understood as a filter that isolates the 'support' component of such data. O'Connor et al. (2010) found that tweet sentiment correlates highly, ranging from r=.77 to r=.81 (p. 128), with opinion poll time series on presidential job approval, adding strong evidence to the positive-sentiment=support hypothesis. This influential finding has since been replicated by Cody et al. (2016), with similar results using the same methodology on new data. Following O'Connor et al's seminal paper, several publications have sought to leverage the linkage between positive sentiment in tweets and approval towards political candidates to forecast election results using tweet volume data mined for positive sentiment. Bermingham and Smeaton (2011) forecast the Irish general election of 2010 using supervised sentiment analysis, as well as benchmarking their analysis with mention data, but predicting parties' vote shares by adding all variables into regression models (fitted to the polls). However, the lowest MAE achieved by these models does not match that of traditional opinion polling, always exceeding 3%. Sang and Bos (2012) follow a similar paradigm with the Dutch Senate elections, whereas their measure of accuracy is at the Senate-Seat-level, where 8 seats out of 75 (p. 59) were falsely assigned to a given party. Furthermore, they stratify their sample by only allowing one tweet per tweeter, and the mention of only one party per tweet. The most methodologically rigorous (and predictively accurate) research in the sentiment-based forecasting domain to date was conducted by Andrea Ceron and colleagues (Ceron et al., 2014, 2015). The authors use a supervised sentiment analysis algorithm based on the work of Hopkins and King (2010). Using this approach, the authors were able to forecast the outcome of the 2012 US presidential election with a MAE of 0.02 % (Ceron et al., 2015, p. 11). Furthermore, they were able to forecast vote shares in crucial swing states more accurately than the average of pre-election opinion polls in 8 out of 12 cases. They achieved these results by filtering tweet collections to only include statements containing a clearly extricable voting intention as well as a *positive* party or candidate mention.

On aggregate, sentiment-based methods have higher predictive accuracy than volume-based methods. However, there is no one perfect sentiment analysis tool for measuring sentiment in relevant tweets. In general, different approaches to sentiment classification can be divided into *dictionary-based* approaches (a matching algorithm looks up target words in an annotated dictionary, retrieves the sentiment score (usually an ordinal scale ranging from -*n* to +*n*) and adds up sentiment scores of all words in the text unit to provide an aggregate sentiment score (see e.g. Thelwall et al., 2012), and supervised/machine learning approaches, where a sample of the target population of text units is annotated for sentiment, and then used to train an algorithm that measures it in a whole collection (Hopkins and King, 2010). On average, the machine-learning approach has proven more valid and reliable, however, it requires significantly more resources, as every new sentiment classification task demands newly labelled data.

Furthermore, there is no clear consensus as to what sentiment analysis for tweets is actually measuring - is it a linguistic classification problem seeking to analyse the intensity and polarity of *language* without analysing what such language signifies in a given context, such as politics or elections? Or is the classification task seeking to extract "political sentiment" or *stance*, in other words what language means in the specifically political domain? This is a further strand of inquiry this thesis adds to, by developing a theoretical framework for political tweets and a novel approach to sentiment extraction and modelling vote shares using them, in 'Understanding Political Sentiment'.

Besides the widely employed approaches introduced above, some researchers have sought to leverage data other than that derived from the content of social media postings (i.e. tweet text, from which volume and sentiment volume are extracted). For instance, Barclay et al. (2015) found that the volume of "Likes" on verified candidates' Facebook pages were a strong predictor of their vote share in the 2014 Indian Lok Sabha elections. However, this finding is not in line with Giglietto's (2012), who forecast the Italian mayoral elections of 2011 with the same methodology - here, no significant correlation with Facebook likes and vote shares was observed. Ceron et al. (2016) note that candidates' follower numbers - which can be seen as the Twitter equivalent of Facebook candidate page likes - should not be understood as indicative of vote shares, but rather of attention to a certain candidate (p. 19). They use the example of Barack Obama and Mitt Romney's follower numbers on Twitter during the 2012 US presidential election: Romney was out-followed by Obama by a factor of 17, while the election result was considerably closer. Lui et al. (2011) found that Google Trends search volume data was also not a good predictor of vote shares in the 2008 and 2010 US presidential and midterm elections, suggesting that these data fall short of Twitter volume data in their predictive accuracy of elections. Franch (2013) combined data from Facebook, YouTube, Google Trends and Twitter with polling data to produce very accurate vote share forecasts using multi-level auto-regressive models. Beauchamp (2017) used features extracted from Twitter volume data, subset by individual tweets' authors' US state locations, to predict and interpolate state-level polling leading up to the 2012 US Presidential election. The author found that Twitter mention data can indeed be successfully used to predict and interpolate polling data, but the predictive accuracy declines the further away the time frame of tweets is from the target poll. Wang et al. (2015) used data obtained from daily vote-intention polls given to users of the Xbox Live video gaming platform to show that when statistically processed in the right way, explicitly non-representative public opinion data can produce election forecasts on par with those derived from representative polls.

2.5 Gaps in the literature and research agenda

In the previous sections, I have outlined the three core areas around which this thesis is centred. First, I described the term *public opinion*, delineated its origins, and how understanding of it has shifted over the years. Then, I outlined how it is commonly measured, namely using survey research methods. Here, I explained the current problems facing the discipline. Then, I introduced the concept of forecasting, and more specifically forecasting election results ahead of time, and described the key academic output from election forecasting. Finally, I reviewed the existing literature on public opinion and elections that employs Twitter data, dividing the literature into two broad categories - those that use tweet keyword volume as a measure for electoral support, and those that use positive sentiment tweet volume.

At this stage, then, the question is - what is missing, and how do I intend to contribute? I argue that there is clearly a lot of evidence on public opinion, how it can be measured, and how it pertains to electoral outcomes. At the same time, I have also illustrated why I believe it to be important for public opinion researchers to go beyond what surveys can offer, especially given the current issues with declining response rates and an emerging trend of survey-based pre-election polling to get the ′mood of the public′ wrong, most notably in the cases of Trump and Brexit. The logical step, then, is to advance the understanding of what analysing social media data can add to our understanding of public opinion, and further, what we can learn about the overall utility of social media data for social science research. However, this comes with numerous problem areas that need to be illuminated in order for the discipline as a whole to even consider catching up with the established prowess of surveys, and further factors beyond (more or less) actionable problem areas remain: I have discussed how platform design and decisions shape digital trace data, and furthermore, the social desirability bias inherent in surveys is likely to be just as relevant when it comes to the potentially endless reach and context collapse-prone world (Marwick and Boyd, 2011) that is one-to-many digital communication. But, in the end, these data exist, in a quantity previously unimaginable for social scientists - and if academics, acting in the public interest are not the ones who spearhead their analysis, it will be profit-driven actors whose incentives are unlikely to align with the public's.

In order to define this research agenda, it is useful to understand where the actual benefits of ′big data′ lie:

> *"Although the increase in the quantity and diversity of data is breathtaking, data alone does not a Big Data revolution make. The progress over the last few decades in analytics that make data actionable is also essential. So Big Data is not mostly about the data."* Gary King in Alvarez (2016)

This understanding is crucial. Big data, or any data for that matter, mean little without interpretation. However, now this interpretation has to happen at scale, thus eclipsing the capacity of human intelligence to efficiently process it. Hence I would add an important caveat/addendum to King's assessment: while it is true that the importance of efficient analytics ('the algorithm') may outweigh the importance of (all) the data, I would argue that it is equally, if not more important, that prior to developing ′the algorithm′ that is more efficient, faster, and smarter than what came before it, there is a strong theoretical foundation that guides and informs the understanding of what it *intends* to do, and what it *might* end up doing. Furthermore, this robust theory requires debate and deliberation, as, once "′the algorithm′ is handed over from the brains of its creator to the machine, it generates data - scores, estimates and classifications - of its own, which will undoubtedly be used by others at a later stage without necessarily having a comprehensive understanding of the theory and rationale underlying the generation of this data. We are currently seeing widespread reporting of examples of data analytics and applied machine learning which reproduce the biases of their creators, leading to horrific consequences for those individuals affected by them (see e.g. Lewis, 2018; Simonite, 2019). In the field of public opinion research using digital trace data, such biases would likely result in misinformation through inaccurate predictions and findings, but the potential for harm in other areas is vastly higher, and must thus be taken thoroughly into account prior to rushing toward deploying new approaches of analytics.

I define the instrumental goal of this thesis as further investigating the feasibility of Twitter-based election forecasts. As stated before, I see elections as unambiguously true outcome quantities[13] to which models can be benchmarked, and iteratively improved. For this purpose, however, it is important to work on the theoretical conceptualisations of the relationship between digitally mediated human behaviour in the form of social media posting, and human political participation in the form of voting. I argue that it is not as simple as merely tallying up mentions of 'Trump' and equating this with votes for Donald Trump. This may work on some occasions - as indeed it does in my tentative demonstration in the introduction of this thesis. But there is no understanding of *why* that is the case. Did one account tweet 'Trump' 7000 times a day? Did leftists tweet their dissatisfaction with Trump? The likeliest answer is that both are true, and the reason that in this case, the 'forecast' got it right is somewhere between coincidence and an incredibly noisy, unreliable instrument.

So, I argue that several things need to be attempted in order to truly evaluate the feasibility of Twitter-based public opinion research, all of which I investigate in this thesis. First, the focus must be shifted from the tweet- to the user-level. Ideally (for a second ignoring bot accounts), every Twitter account belongs to a human, with (political) opinions and preferences, some of which they express on the platform. In several cases, they express these opinions and preferences over time, and a pattern may emerge. Second, we need to know who those users are, and what makes us able to sort them into groups. I am referring to the computational estimation of socio-demographic and other individual-level attributes of Twitter users, a fruitful means of better matching Twitter data samples with target populations. Third, and finally for this PhD thesis, there stands the question of how we can best transpose the content of the digital trace data these users produce - in the form of tweets and associated metadata - onto the conceptual frame of public opinion. What can be done to overcome the limitations of keyword mention-volume or positive sentiment-volume designs for estimating public opinion from tweets, to enhance them and to add to them? I address all of these questions from different angles across and throughout the three papers presented in this thesis. While I outline the paper's aims and research designs in detail in the next chapter, in essence, I address the unit of analysis question in all three papers, I contribute a new method for geo-locating Twitter users, i.e. an approach to estimating user-level socio-demographics, in 'Finding Friends', and I delve further into how best to extract individual-level political preferences from tweets in both 'Understanding Political Sentiment' and 'Listening in on the noise'.

[13] This excludes of course the myriad numbers of elections in human history where evidence of cheating, forgery and tampering is ample, and thus restricts this framing to developed, mature democracies

Chapter 3

Research Design

Having discussed why knowledge of public opinion is normatively desirable for the functioning of democratic political systems, and how established approaches to measuring public opinion are becoming increasingly more costly and difficult to conduct, this thesis sets out to add knowledge regarding **the research questions of *if* and *how* user and tweet-level data collected from Twitter can add to the understanding of public opinion as it pertains to elections**. In other words, this thesis seeks to understand whether, and if so, how instances of political speech on Twitter can be mapped onto what people want, politically, in the offline world. There is a strong intuitive, theoretical and indeed empirical case for why this should be the case: social media data, and especially tweets are plentiful and oftentimes concerned with politics. Further, individual data points (i.e. tweets and their associated metadata) provide rich, qualitative behavioral information about their authors. In many ways, the literature review preceding this chapter provides multiple justifications for why the answer to the *if*-question should be considered trending towards 'yes': the amount of existing research in the area of mapping digital trace data to public opinion, even if not consistently successful, shows that there exist avenues for sound and insightful research in this area. These further offer clear opportunities for more exploration and different approaches. Hence, much of this thesis focuses on addressing the *how*-question by developing and benchmarking new methodological approaches for making such research more reliable, reproducible and broadly applicable. In doing so, and by providing not only novel methods and open source code for others to re-use and adapt, this thesis also adds empirical evidence to the field, with the aim of making the *if*-question; the question of the feasibility of Twitter-based public opinion research; easier to answer in the affirmative.

This agenda of a diverse methodological and empirical contribution to the field is approached from different perspectives in three self-contained research papers within this thesis. First, I introduce, validate and demonstrate a novel algorithm and software pipeline for geo-locating Twitter users to their home locations. This allows for a geographic disaggregation of where politically salient tweets originate from, and whether they should be considered relevant to any kind of model that measures public opinion. For instance, users from Indonesia may tweet their opinions on US politics. However, these users are not entitled to a vote in the US - and thus such tweets should not feed into further analyses regarding Americans' political views. Therefore, this paper provides a solution for the problem of selecting the correct data from a desired sampling frame for Twitter-based public opinion research, as without it the parsed opinions of Indonesian Twitter users posting about US politics will be hard to exclude from relevant analyses conceptually limited to US users. Hence, this paper and the method presented therein provide a foundational step for furthering the understanding and methodological toolkit of *how* Twitter data can be analysed for signals pertaining to offline public opinion. This is further underlined by the fact that the methodological contribution from this paper feeds into the empirical analyses presented within the remaining papers.

As such, the second paper, 'Understanding Political Sentiment', moves away from the foundational methodological question of accurate sampling, toward modelling public opinion from tweets. Specifically, the focus is on sentiment analysis, a promising and widely employed method for capturing Twitter-based public opinion in the existing literature, especially when applied in the election forecasting scenario. This paper expands widely on existing applications of the paradigm, by rethinking the operationalisation of 'sentiment' when used in the context of tweets and public opinion, and by benchmarking and comparing multiple ways of aggregating measured and estimated user and tweet-level sentiment to geographically defined areas in the electoral vote share mapping scenario. This way, this paper contributes significantly to the *if*-question of Twitter-based public opinion research, as it provides strong empirical evidence showing that the sentiment analysis paradigm *can* work, and when one should expect it to be accurate. Further, it provides a strong contribution to answering the *how*-question, as innovative methodological adaptations are presented and compared to existing approaches, while also comparatively presenting their relative efficacy for the pursuit.

The third paper again takes geo-located users as its starting point and then models users' voting intentions in the 2018 US midterms by applying Machine Learning and distant supervision on users' tweets on relevant topics *related* to the 2018 election. This paper further showcases the importance of sampling decisions in Twitter-based research, as it applies identical modelling approaches on a purposive sample of likely voters and a random sample of Twitter users. This shows that, when aggregated to target offline populations of interest, e.g. users/voters from California or users/voters from Texas, results differ widely, both when compared to one another, and when compared in aggregated form to state-level vote share tallies in the 2018 midterms, the key instrumental paradigm employed in this thesis for assessing the external (i.e. offline) validity of Twitter-based public opinion modelling. So, this paper again furthers the understanding of best practice for Twitter-based public opinion research by highlighting the effects sample selection criteria can have on eventual findings, and deducing recommendations for future research. It also innovatively contributes to the *how*-question by outlining a method that overcomes the complexities associated with identifying pertinent users/tweets and the transposition of their published political content into political and public opinion-salient categories in the offline world, even if content directly indicative of users' political preferences does not exist. Furthermore, this paper offers the strongest evidence in this thesis that offline public opinion can indeed be measured from tweets with its distant supervision vote-choice modelling approach. This further adds to the basket of arguments in favour of a 'yes' answer to the *if*-question.

Overall, this thesis contains a cumulative contribution to the field in both foundational/conceptual regards (*if*) and methodological/practical aspects (*how*). This is developed across three separate papers, building on the thorough review of related and existing research in the preceding chapters. Each of these papers addresses a different sub-component of the overall puzzle of how tweets can be used to measure public opinion, in a fashion where subsequent, self-contained research builds on methodological and empirical advances outlined in previous work, most notably with the application of my geo-locating algorithm as a foundation for reliable sample composition in papers two and three.

In the following sections of this chapter, I first provide a more detailed definition of this thesis' two core research questions, while also indicating how this thesis addresses them. After summarising the resulting cumulative contribution of the thesis, I outline the overall design decisions taken in this thesis, namely case selection, methodological imports, data collection, ethical implications and considerations of this body of work, as well as external implications to this research, its feasibility, reproducibility and future applicability.

The *IF*-question: Can public opinion be reliably extracted from Twitter?

I now describe the core questions guiding the inquiry in this research project in more detail, starting with the epistemological question at the heart of the thesis: Can public opinion be reliably extracted from Twitter?

Assuming there was no way of measuring public opinion whatsoever - surveys and polls had never been invented, and perhaps even elections and censuses did not exist[1]. The question to ask, then, is whether data obtained from Twitter would provide information relevant and informative to a conceptualisation of public opinion outlined above, and would add to knowledge of public opinion beyond the baseline (which in this case, would be *no knowledge*)? The answer has to be yes. Especially in the context of gauging the reaction to (political) events as they happen, Twitter is a useful tool. On the one hand, it allows its users and those interested in the 'mood of the minute' to identify which positions are popular and shared by influential *opinion leaders*. On the other hand, it provides a *public* forum for individuals to share their views on any given topic. Given the sheer size and scope of Twitter, and social media sites in general, the complexity and granularity of individual-level viewpoints that should conceptually be gleanable from Twitter collections is hugely promising. Nonetheless, some individuals' views may never be present on Twitter, because these people are simply not present on the platform. However, hard or impossible-to-reach populations exist in most any type of social science research, and I argue that the importance of converting these from 'unknown unknowns' to 'known unknowns' forms a natural component of any area of maturing scientific inquiry, and should in itself not warrant a dismissal of the overall feasibility of Twitter-based public opinion research. For this purpose, I understand the goal of improving user-level characteristics, and comparing them in aggregated form to target offline populations as a crucial component of strengthening the case for the feasibility of Twitter-based public opinion research, which is why this thesis devotes considerable space to it.

As outlined earlier, there have been multiple examples of research that uses mention volume of certain target keywords on Twitter as a proxy for public opinion. How reliable would information produced like this be, and further, how (externally) valid can this information be, given it comes from a narrow, self-selected pool of subjects, with varying, *a priori* unknown

[1] This would be the case, for example in non-democratic regimes like Saudi Arabia, where, incidentally, the per-capita Twitter adoption rate is very high compared to western countries, and participation on the network is equated by some as 'the town square' (Hubbard, 2019)

propensities to contribute pertinent content? Clearly, there are several factors affecting reliability and validity. First off, there is the frequent mis-mapping of Twitter's population versus target populations, as well as the lack of information about users' demographics. Addressing these shortcomings forms the key contribution of my first paper and geo-locating algorithm. While this does not cover the entirety of Twitter users' demographics, it addresses a major indicator of inter-individual variability of political behaviour.

Further, having conceded that certain individuals' views may never be represented through Twitter (but also stressed that this is not a necessary condition for its usefulness, given knowledge of its population and target populations), one may still argue that Twitter membership does not entail an obligation to tweet, and indeed certain people may never explicitly share their views on issue *i*. One may feasibly use this argument to justify that this self-selection bias inherently skews insights derived from the platform and cannot adequately be adjusted to form a more balanced picture of *who wants what*, because only those who explicitly state what they want are heard. This would support the position that reliable and useful Twitter-based public opinion research may *not* be feasible. For this purpose, I argue that another vital strand of inquiry that must be pursued in order to strengthen the case for Twitter; the *if*-question; is that of parsing out (politically) pertinent information from entire samples of users/tweets, rather than just those that tweet explicitly about a target issue of interest which can be matched with pre-defined keywords.

In my third paper, "Listening in on the noise", I address this issue by using users' related content - such as tweets mentioning keywords related to pertinent news stories, candidates or other issues related to an upcoming election - to estimate their voting preference. In other words, I am able to construct a measure of a user's view on *i* even if they never tweet about *i*. Of course, some users simply never tweet - but for anyone that does, this method significantly expands upon the ways in which their tweeting can be parsed into useful quantitative and qualitative information which can inform their individual-level (political) views. Crucially, this expands considerably upon the scope and breadth of existing Twitter-based public opinion analyses, and therefore alleviates those concerns that suggest that the pursuit overall falls short because we can only measure those peoples' opinions who actively and explicitly share them with keyword-retrievable language.

In the preceding paragraphs, I have outlined the core questions addressed in this thesis which deal with the question of the feasibility of the pursuit of Twitter-based public opinion research; with the raison d'être of this line of inquiry as a whole. At this point, it may be clear to the reader that I firmly believe in its justification and feasibility, however, this thesis emphasises its contribution to a more rigorous and mature practice by addressing several key shortcomings of existing research in this vein. These are, firstly, the perceived lacking of overlap and adjustability between Twitter and target offline populations, and further the supposed inability to measure the (political) views of those who do not tweet immediately, explicitly relevant content. Through my methodological innovations in these areas and fruitful empirical applications thereof, I add considerable evidence in support of an increasingly more confident affirmative answer to the *if*-question.

The *HOW*-question: How can public opinion reliably be measured from Twitter data?

Studying public opinion through the lens of Twitter data does not come with a comprehensive, well-established methodological toolkit[2], given the novelty of the field, and the lack of consensus on some of its most important methodological parameters. Hence, this thesis follows what may be considered a *research design philosophy* of addressing some of the many under-developed areas of the field in an *exploratory* framework. In other words, the goal was to conceptualise some of the many unanswered questions from the perspective of a social scientist, as, crucially, this thesis is situated within the emerging field of *computational **social science***, a discipline which applies methodological approaches imported from computer science and data science to address social science questions.

This thesis aims to re-focus the pursuit of Twitter-based public opinion research around core principles of social science: rigorous definition of desired outcome quantities and increased dedication of resources to operationalisation versus 'the best possible model', awareness and identification of sources of potential bias, and extrapolating from samples instead of 'zooming in' from pseudo-populations (see the "*n=all* fallacy" (Jungherr, 2017)).

Having outlined the crucial factors which warrant inspection and discussion when considering the feasibility of Twitter-derived measures of public opinion, I now move on to discussing the methodological gaps in the field which this thesis addresses; in other words: the *how*-question of best practice in Twitter-based public opinion research. It is important to note that this is not a comprehensive list of *all* existing gaps in the field of Twitter-based public opinion research. Rather, it is what I see as *the most important* gaps, the investigation of which will contribute most to both assessing the feasibility of the endeavour as a whole, and providing most value to the field going forward. I have identified the following areas this thesis contributes to:

1. Best practice for *sampling*, i.e. how to define sampling frames and how to produce the best possible sample with external validity in mind

2. Exploring how *hand-annotated data* can best be generated and analysed for the purpose of parsing public opinion from tweets, and further, how hand-annotated scores can be propagated to larger samples of tweets or users, regardless of a given user having necessarily tweeted something captured by a given *a priori* defined keyword

3. Assessing the impact the unit of analysis can have on research findings (i.e. tweet-level versus user-level)

4. Developing methods of *estimating/classifying user-level socio-demographic attributes*, in order to make more informed sampling decisions.

The first paper, 'Finding Friends', revolves predominantly around developing, testing and applying a scaleable and reproducible method for assigning estimates of users' home locations to user-level metadata. At its core, this addresses the crucial need for reliable

[2] A possible exception is (Klašnja et al., 2017), which provides a comprehensive overview of the existing challenges and opportunities that exist in the field of public opinion measurement from social media. However, this paper, as useful as it is, focuses mostly on under- and un-explored areas of the field, meaning that it cannot be considered a "guide"

information on Twitter users' socio-demographics. Being able to discard users who are not entitled to a vote in a given geographical entity, regardless of whether they have posted their opinion on its politics or not - allows for more fine-grained and robust research. This is more broadly applicable to any research where one may want to restrict ones' sampling frame to include only data originating from a defined geographic entity. Users' home locations are classified using a hybrid methodology, either matching their self-reported location string to census records or leveraging their reciprocal friend-follower network (RFF) to estimate their own location. This methodological advance feeds into both of the following papers in this thesis, and has further been used as a foundation for sample generation in colleagues' work (Barberá et al., 2019; Brown et al., 2020; DeVerna et al., 2021). Besides its innovations in methods for estimating user-level attributes, this paper also adds evidence to the importance of the unit of analysis in Twitter-based public opinion research. In applying the geo-locating method to a pertinent example, namely the mentioning behaviour towards Democratic party presidential candidates at the US state-level, I show that a user-level analysis yields vastly different, and in this case, more externally valid measures of public opinion than a tweet-level analysis. I suggest that this originates from 'prolific partisans' - individuals who frequently tweet about the same candidate and thus artificially amplify their overall mention numbers on Twitter as a whole, and within states specifically.

'Understanding Political Sentiment' adds evidence to several of the issue areas selected for investigation. Most importantly, the paper explores the question of the role of hand-annotated data in Twitter-based public opinion research. As is the case with much of contemporary text-based data analytics and natural language processing (NLP), computers and algorithms are actually *not* 'smart' enough to distill meaning from free-form text without help from humans who read the text, interpret it and annotate it to reflect some desired outcome quantity. This is no different when it comes to tweets: regardless of whether one may want to know the linguistic sentiment of a tweet, or whether one may want to classify a user's political ideology, humans are required to interpret and evaluate this. Existing research in this field has however not strayed far from the traditional sentiment analysis paradigm: tweets are classified for their linguistic sentiment, and machine learning methods are used to broaden the classification from a small training sample to a larger population. First, I develop a granular annotation framework for ordinal *political* - not linguistic - sentiment in tweets. This is applied on tweets by unique users from different states. I then benchmark how data labelled in this manner can be processed and modelled to measure offline public opinion in the election vote share estimation paradigm. Besides applying established approaches of equating tweets about political candidates annotated or machine-classified as positive to signal electoral support for that candidate, I describe a novel approach to modelling individual-level vote preferences from labelled data, and further apply machine learning regressors with the goal of maximising the granularity of machine-generated scores when propagating these ordinal scores to larger samples of un-annotated data.

'Understanding Political Sentiment' further investigates the impact of the unit of analysis, by comparing tweet-level models to user-level models in the vote share aggregation application of measuring public opinion from tweets. By only changing one component in a given model and applying this across different samples, this adds strong evidence regarding the respective applicability of a given approach. Furthermore, this paper critically investigates the notion that more data results in better findings, instead pursuing the idea that sampling can be a very useful tool when working with large-n collections of tweets. Models are not only applied on differently conceptually defined (by geographical inclusion criteria) samples, but performance is also compared between comparatively small (n=1000) samples of purely hand-annotated data versus significantly larger samples of machine-propagated data.

Finally, 'Listening in on the noise' addresses several of the problem areas identified above. Firstly, the focus is on sampling, and more specifically the impact of purposive versus random sampling; namely comparing keyword-derived tweet collections with specific criteria versus randomly generated samples of tweets or users. Given the fact that this research relies on 'big' data - there are around 500 million new tweets, every day - sampling plays a different role than it does in traditional social science research, where most empirical data is sampled by necessity. Hence, the question arises: why sample, if there is so much data to choose from, and we have the computing capacity to analyse all of it? Indeed, this has been the dominant approach of previous publications in this field. However, there is a strong argument (outlined by Jungherr (2017)) that it is impossible to know whether a 'big data' collection is a population or rather a (biased) sample, as it is impossible to know *a priori* whether inclusion criteria are comprehensively broad or narrow; or, specific to the case of Twitter data, if Twitter's API is actually returning the population of tweets that exist for the defined inclusion criteria, or rather a sample - with unknown sampling parameters. I investigate this by running identical analyses on two distinct samples, one purposively created to contain *previous voters*, determined through user-level tweet content, and one random, generated with a random-number generator matching user-ids. Going by external validity in the vote share prediction paradigm, this paper finds that the random approach is considerably more fit for purpose.

Secondly, this paper further adds to knowledge on use-cases for hand-annotated data in Twitter-based public opinion research. While 'Understanding Political Sentiment' sets forth to broaden the application of sentiment analysis, here users are coded in order to reflect their explicitly observable voting intention. This then allows me to select all those users for whom such a value was assignable, and then use their corpus of published tweets to train a machine learning classifier which can probabilistically classify any given users' voting intention. This provides a theoretically sound, and, as the findings indicate, empirically highly useful method for estimating individual-level political preferences for an expanded set of users, namely those who explicitly share whom they intend to vote for and those who do not.

A Cumulative Contribution

As stated earlier, the thesis follows an exploratory research design philosophy. Given the novelty of the field, and the evolving nature of its underlying behavioral patterns distilled into analysable data, I argue that this is the fruitful pathway toward improving the state of the art in Twitter-based public opinion research, but also towards an honest investigation of the feasibility and usefulness of poll-like evidence derived from digital trace data. Hence, the core contribution that results from this thesis should not be seen as one method or one finding to rule them all, but rather as a collection of methods and evidence which add to the growing body of evidence and guidance on best practice, as well as function as justification for why researchers may want to study public opinion using data gathered from Twitter. To reiterate, the key areas this thesis contributes to are manifold: I investigate sampling of digital trace data, both in contrast to the "big data" paradigm where sampling is seen as something that reduces the quality of findings, and as a question of keyword-derived samples versus random samples. Further, this thesis features a multi-faceted inquiry into what a shift towards user-level analyses can bring to the field as opposed to commonly used tweet-level analyses. I showcase several approaches to hand-labelling tweets, and further introduce new methods for analysing such data with public opinion in mind. Finally, I present a method for classifying users' home locations, so as to allow geographically disaggregated analyses of Twitter data. I suggest that all of these contributions are valuable beyond the application of Twitter-based public opinion research, but can rather prove useful to computational social scientists working with Twitter data as a whole.

This cumulative contribution addresses several of the recommendations made by (Gayo-Avello, 2013, p. 671) in his influential review of the state of the field, namely estimation/classification of users' demographic attributes, expanding the approach towards sentiment analysis, and also addressing self-selection bias, by computing voting intention for users who did *not* tweet something immediately identifiable as relevant through keyword filtering.

The thesis taps into a hierarchy of problems that exist with Twitter-based public opinion research. While there was not a full research plan at the outset of the project, there was a clear focus on wanting to solve some of the problems existing in this research area, keeping in mind their relative urgency, as interpreted through the lens of the state of the field. Throughout the research, the identification of a hierarchy of problems resulted in the content, sequence and configuration of this thesis. This has led to a range of methodological and empirical contributions, all of which can be useful for future research studying public opinion through the lens of Twitter and other digital trace data. I will revisit the question of the lasting cumulative contribution, as well as its role in future research in the discussion chapter of this thesis.

Case Selection

In order to apply the methodological contributions introduced in this thesis, it is first necessary to select empirical cases on which to test them. Throughout this thesis, I test methodological approaches for measuring public opinion from tweets in the context of elections, as they provide an unambiguously true outcome quantity to which model estimates can be compared, and their efficacy and usefulness therefore be evaluated. In 'Finding Friends', I measure attention and support to declared and presumed electoral candidates. In 'Understanding Political Sentiment' and 'Listening in on the noise' I evaluate given methodological approaches and models by applying aggregating estimates in the election forecasting / vote share prediction paradigm. As stated earlier, this is to be understood as instrumental to the overall goal of this thesis, which is to advance the field of Twitter-based public opinion research, rather than as a pursuit for the best possible election forecast for its own sake.

Given that my substantive expertise and research interest is centred on U.S. electoral politics, I chose to select elections taking place in the United States as the context in which to apply my methodological contributions. For 'Finding Friends', this is the 2020 Democratic party presidential primary, whereby I measure geo-spatial variation of attention and support towards different candidates. For 'Understanding Political Sentiment', I aggregate 'political sentiment' scores using different methods in order to predict candidate-level vote shares in the 2016 US primaries in New Hampshire, South Carolina and Massachusetts. I specifically selected these states as they provide exceedingly different socio-demographic attributes and population distributions, while featuring the same candidates and highly similar electoral rules. For 'Listening in on the noise', I aggregated individual-level estimated vote choice variables to predict party-level vote share percentages for Congressional elections in California, Texas, New York and Florida, as well as the national popular vote. This case selection again allows me to hold several factors constant (electoral rules, parties) while factors such as socio-demographics, population distribution and party identification vary across different states.

Overall, I gave considerable importance to the possibility of testing approaches in different locales at the same time, as this would provide a solid foundation of evidence upon which to evaluate the relative efficacy and usefulness of a given method of measuring public opinion from tweets. Furthermore, I argue that the focus on smaller, less high-profile elections (as opposed to a U.S. presidential election) offers a more nuanced look at how the methodological contributions in this thesis work 'in the wild', in more complex configurations than a two-candidate, winner-takes all horse-race. However, it is also important to note that an empirical examination of a U.S. presidential election would have been highly interesting and useful, but did not fit into the timeline of my PhD research: the 2016 election came too early, while 2020 came too late.

Methods

None of the methods employed in this thesis are novel in the sense that algorithms, models and statistical approaches have been widely used, in some cases over decades. The novel aspects described in this thesis are their application on data generated in a novel fashion, paired with specific pre-processing steps, within a soundly theorised conceptual framework, thus allowing a new framing and combination of existing tools to allow for new approaches. Hence, this thesis provides considerable methodological innovation.

Many of the methods applied in this thesis draw from the fields of machine learning, natural language processing and applied statistics. This is particularly pronounced for methods used to regress or classify variables. In 'Understanding political sentiment', I employ machine learning regressors, namely Linear Regression, Logistic Regression (one versus rest), Multinomial Logistic Regression, Random Forests and Ordered Logistic Regression as a means to propagate measures of candidate support (political sentiment scores) to users where such a score was not hand-annotated. In 'Listening in on the noise', I employ the Naive Bayes classifier on hand-annotated tweet text to estimate the probability of a given user voting for a given party. I further use Logistic Regression in an ensemble model of vote choice classification that incorporates further user-level metadata. Throughout the thesis, I further use established methods which gauge the performance of a given model, regressor or classifier, namely the framework of precision, recall and accuracy, as well as mean absolute error. I also rely on descriptive statistical methods, such as Pearson correlations, as well as methods used to describe text, such as Cosine similarity and textual distance.

All statistical and computational analyses performed in the framework of this thesis were authored by myself (and in the case of 'Finding Friends', my co-author, Andreu Casas) using the open source programming languages R (Version 3.4) and Python (Version 3.7). All inferential, predictive, classification, regression and descriptive methods described in this section were applied using built-in or external modules (cited within papers) for the respective languages. As none of these modules provide warranties for their implementations of a given algorithm or classifier being 100% accurate, I also cannot provide such a warranty. However, given the open source nature of the software packages, and the ample participation of the community in improving them, I am highly confident of their accuracy. All code used in the production of this thesis is available on GitHub (see links in the front matter of the thesis).

Data Collection

The thesis is not only methodological, but also empirical and data-driven in nature. The core data used for this thesis are tweets, and user profile information collected from Twitter using its publicly available and free APIs (Application Programming Interfaces). It is comparatively simple and cheap to collect data from these APIs - the only requirements are a verified developer account with Twitter and an internet connection. Then, by generating and registering an application, Twitter provides the user with credentials ('keys'), which are then used to authenticate the requests sent to Twitter's API. Such requests can range from collecting data from Twitter's stream of tweets as they happen (i.e. live), to collecting historical tweets or user-level metadata/profile information. Requests for data retrieval are scope and rate-limited, meaning that specific requests can only be performed up to specified quotas, or be repeated a limited number of times before the service times out, and no longer returns the requested data.

Baseline tweets forming the sampling frames for both 'Understanding Political Sentiment' and 'Listening in on the noise' were taken from New York University's Social Media and Political Participation Lab's (SMaPP) on-going Twitter collections pertaining to the 2016 US election. These large-scale collections contain all relevant tweets containing a large and comprehensive list of keywords pertaining to the elections[3]. For 'Understanding Political Sentiment', these tweets were filtered for the pertinent date ranges - 3 weeks leading up to each of the target primary elections. Then, the individual Twitter users contained therein were geo-located, discarding users who were not from target states. Remaining tweets were then used for analysis.

Hand-annotation in 'Understanding Political Sentiment' and 'Listening in on the noise' was performed by me. For 'Listening in on the noise', baseline tweets were collated from the aforementioned 2016 presidential election collections. The baseline for the purposive sample were previous voters, i.e. users in the collection who tweeted keywords indicative of having voted (see the methodological outline in the paper). The baseline for the random sample was a set of random Twitter users, previously generated by researchers in the SMaPP lab by generating random numbers in line with Twitter's user-id structure, and testing if the randomly generated number is a user, including it if so. Unique users from the baseline tweets were geo-located, and subset for the inclusion criteria (home location in the United States). Then, I collected all remaining users' tweet histories, as well as incoming new tweets for the time-frame of interest using Twitter's *getUserTimeline* API endpoint, implemented using the rtweet package for R (Kearney, 2018).

[3] While I could have also performed these tweet collections myself, the problem of large-scale data storage (a day of tweets pertaining to the 2016 US election often exceeds several gigabytes), it made sense to avail myself of this existing resource in my capacity as assistant research scientist at the lab.

Tweets used for both method validation and empirical application in 'Finding Friends' were collected by NYU's SMaPP Lab. This includes the aforementioned randomly selected Twitter users collection, as well as a comprehensive collection of all instances of mentioning a declared or presumed candidate for the 2020 Democratic presidential primary over the course of 2019. This paper also avails itself of data from a linked survey conducted by YouGov, commissioned by the SMaPP Lab in 2016. Home locations classified using the 'Finding Friends' pipeline were produced using data gathered from Twitter's *getUser*, *getFriends* and *getFollowers* API endpoints, as well as geonames.org's *search* API endpoint. These requests were implemented using a combination of the rtweet package for R (Kearney, 2018), the tweepy package (Roesslein, 2020) for Python and a custom wrapper for the geonames API written in R by me.

All retrieved data are stored on password-protected and encrypted network and local drives. Access is only available to myself, or research collaborators. No individual-level Twitter data is published in this research, or shared with any publication of the thesis, in accordance with Twitter's privacy policy.

Ethical Considerations

Data collected from Twitter are public on the web, and freely and legally available for collection by anyone through Twitter's public APIs. However, it is also important to note that it would be unreasonable to assume that the majority of individuals whose tweets and metadata were incorporated in this research project in some form are aware that this is the case, and may have considered the possibility of their tweets being used in this fashion. Hence, it is important to consider individual users' privacy when using individual-level data for whose use informed consent cannot feasibly be granted. For this purpose, I follow a number of guidelines in order to ensure minimal risk to users whose data was used:

- No reported research findings contain verbatim tweet text

- No reported research findings include any identifiable individual-level characteristics, making it impossible for anyone to conclude users' identities based on the content of this thesis

- No individual-level data collected from Twitter, or computed from it, is shared with anyone beyond immediate collaborators involved in this research

- Data is stored securely and encrypted, with access only for myself and immediate collaborators

- IRB/ethics clearance was provided for this research project by the School of Social Sciences at the University of Manchester as well as the Faculty of Arts and Sciences at New York University.

Overall, I argue that the measures outlined above prioritise the rights of individuals whose data have been used in this research, and present a strong framework for the minimisation of any potential ethical conflicts arising from the research.

This research project was supported by the Economic and Social Research Council of the UK, the School of Social Sciences at the University of Manchester and the Data Science

Institute at the University of Manchester. I have no conflicts of interest to disclose resulting from this funding.

Constraints and External Limitations

While this thesis offers a comprehensive investigation into epistemological and methodological factors associated with the feasibility of Twitter-based public opinion research, there are certain relevant aspects which the thesis does not contribute to. Furthermore, it is important to note that the enduring correctness and reproducibility of many, if not all findings and conclusions drawn from this thesis are subject to the decisions of an opaque corporation (Twitter Inc.) not subject to democratic control. Thus it is possible that parts of this thesis may no longer be applicable in years to come.

Most importantly, this thesis does not contribute to questions of the psychology of individual-level one-to-many and many-to-many communications on Twitter, and social media sites in general. Who tweets, who does not tweet, when do which types of individuals tweet, what will their tweets (not) contain given n, and why. Moreover, there exist the questions of which kinds of individuals tweet which kinds of (political) content, when, and, why; and which types of individuals do not? Furthermore, it is clearly useful to study what is observable on Twitter (as it pertains to public opinion) - but what about what is *not* observable? Many users *see* political content on Twitter, and it may influence their political views, and thus public opinion in some facet, regardless of whether they ever tweet about it or not. Knowing their (political) views, and furthermore, knowing *why* they are *not* sharing them, and identifying regularities in the patterns that lead them to share (political) content should contribute to a well-grounded understanding of public opinion derived from Twitter, and more generally, to an understanding of human behaviour mediated by digital platforms. I argue that I adequately side-step the direct need for pursuing this line of inquiry by developing a method for enhancing the ability to parse out signals pertaining to public opinion for users who do not tweet anything explicitly related to a given public opinion-salient issue through my application of vote choice classification assisted by distant supervision in 'Listening in on the noise'. Furthermore, it is important to state that such questions were never within the remit of this thesis; they are not questions that are easily answerable for someone who is trained in political science and (computational) social science, like this author. Rather, they are questions for psychologists, neuroscientists and sociologists. I am however confident that the body of work presented within this thesis can be of use to scholars setting out to investigate such questions, and that the knowledge gleaned herein can enhance both their conceptual problem definition and empirical approach to such complex and important questions.

In regard to this thesis' enduring accuracy/correctness and reproducibility, it is important to note that Twitter is a product supplied for profit by a corporation, whose ultimate goal is to maximise shareholder value. None of Twitter's platform design decisions (which occur frequently, and often without warning) are subject to democratic approval of either regulators or its user-base. It is entirely feasible, that Twitter may suddenly decide to restrict access to data generated by its users to paying customers, to completely cease publicly sharing its data, or, more drastically, that the company goes bankrupt and the entire platform disappears or is purchased by a competitor. So, while a typical user sees a timeline of tweets from other users that they follow and has the option of sharing their own content to the site, this excludes a range of complex, and ever-evolving design decisions the company implements in the pursuit of profit maximisation. This may include, for example, reducing/increasing the relative exposure of political content that certain users see (as,. perhaps, certain users spend less time on the platform if they are exposed to 'too much' political content), or re-organising the sequence of where tweets appear on users' timelines. While this situation significantly complicates the pursuit of understanding 'how Twitter works', and thus also obfuscates the answer to the above-stated IF-question, I argue that dwelling too much on these factors is not expedient for researchers studying Twitter: access to data may well be shut off come tomorrow, so research has to take the platform as is, and work with what is available.

Chapter 4

Finding Friends: A hybrid approach to geo-locating Twitter users

Abstract

A rapidly growing body of research in political science uses data from social media services such as Twitter to study political behavior, including protest mobilization, opinion formation, ideological polarization, and agenda setting. A key critique of this type of research is that the underlying demographics and characteristics of the population of Twitter users are mostly unknown, as are the characteristics of any sample of Twitter users. This makes drawing population inferences, as well as the analysis of particular subgroups on the platform, extremely challenging.

In this paper we present a new method for estimating the location of Twitter users. Many users provide location information in their user-supplied meta-data sufficient to place them in a municipality; but for those who do not, we use information from their "reciprocal follower-friend network" (RFF) to estimate the country, region, and local municipality. The underlying theoretical assumption of this approach is that Twitter networks are geographically clustered, and that we can identify where a user is located based on the sets of known locations of their follower-friend network, thereby accurately identifying the location of the majority of users.

Besides introducing this method and sharing an open-source software package enabling anyone to apply it, we also outline an approach for estimating the certainty of any given location classification. Further, we provide an example research application of this geo-locating method on users who tweeted (n=2.2m) about Democratic party presidential candidates in the first half of 2019. We find that the geo-spatial variation of electorally salient conversation on Twitter closely aligns with both *a priori* expectations and differently obtained measurements (e.g. opinion polls).

4.1 Introduction

Twitter users voluntarily post several hundred million tweets every day. The vast majority of these tweets are freely accessible on the web and retrievable through Twitter's Application Programming Interface (API), thus offering researchers an unprecedented window into public opinion. However, unlike with survey respondents, little information about the typical Twitter user is available by default. Besides a display name (@-handle) and an API-retrievable unique user-id, Twitter's default user profile does not feature any required fields which provide further information on a given user. Optional fields, such as the user's biography (bio) and location are free-form text fields only constrained by the maximum number of characters users can enter.

Nonetheless, users create varying amounts of data which are of interest to social scientists, and political scientists/public opinion researchers in particular. Besides the text or media contained in users' tweets, Twitter's API is also home to a giant amount of user- and tweet-level metadata, which can further be highly valuable for researchers seeking to learn more about social processes. Using such metadata and primary tweet data, it is possible to estimate various latent characteristics of Twitter users, such as age (Nguyen et al., 2013; Sloan et al., 2015), gender (Mislove et al., 2011; Mullen, 2018), education (Culotta et al., 2015) and political affinity / partisan affiliation (Barberá, 2015).

Adding to this literature on estimating user-level characteristics of Twitter users, we outline a reproducible method for estimating Twitter users' geographic home location. We are motivated by two reasons. First, *location is politically relevant*. Location - be it the urban-rural or center-periphery divide - is an important explanatory variable for variation observed when studying political phenomena (see e.g. Johnston et al., 1988; Scala and Johnson, 2017; McKee and Teigen, 2009). This is particularly pertinent when it comes to elections: most democracies hold elections at varying geographic political levels (e. g. state or local elections), making a regionally dis-aggregated analysis of public opinion necessary. However, voters' locations are also highly relevant when it comes to country-wide elections: In the UK's 2016 referendum on leaving the European Union, voters' locations were a very strong predictor of vote choice - even if this variance is likely explained by different underlying variables (Barr, 2016). This means that, in scenarios with unavailable data on causal variables, such as individuals' education or socio-economic status, peoples' home location can help inform a prediction on what shape their political behaviour will take. However, public opinion is not only relevant for the purpose of studying elections: public opinion towards political actors and institutions, as well as towards specific policies influences political outcomes, and *varies regionally*. When China announces tariffs on soybeans from the United States, public opinion towards this policy is likely to differ for individuals from Iowa versus individuals from Alaska. Second, knowing characteristics of individual Twitter users allows researchers to adjust samples in order to reflect desired population measures. While this data pre-processing step may seem intuitive and necessary for most any empirical social science research project employing Twitter data, the practice is far from widespread; precisely because generating reliable estimates of unobserved characteristics of importance is a highly complex task.

Geo-located user-level Twitter data is not only relevant to public opinion research. When studied using Twitter data, fields such as event detection, risk and threat assessment (see e.g. Weng and Lee, 2011), epidemiology (see e.g. Signorini et al., 2011) all hinge on knowledge of the geographic source of tweets in order to reliably process input data. And, while there is an obvious shortcut to achieving this goal - restricting one's sampling frame to *just* GPS-tagged tweets or *just* parsed profile location - the vast majority of tweets (>99%) do not contain any GPS-derived location information ((Sloan et al., 2013, p.7) or (Graham et al., 2014, p.570)), and self-supplied profile locations translate to real, human-readable locations for fewer than half of any given sample of users, as we have found in large-scale explorations of various types of Twitter samples in the foundational work conducted in the context of this paper. Furthermore, evidence suggests that users who *do* geo-tag their tweets differ significantly from those who do not in several key socio-demographic and behavioural attributes (Graham et al., 2014, p.570). Hence, it is necessary to leverage other available user-level metadata to classify Twitter users' locations.

Previous groundbreaking research has used tweet content, tweet and user-level metadata or network data to produce user-level location estimates. However, several shortcomings are found throughout the existing literature yet to be addressed. First, most studies use geo-tagged tweets as ground truth data for training and testing new geo-location algorithms, but as stated above, the extremely rare activity of geo-tagging tweets is likely to be an indicator of atypical users, and furthermore, there are good reasons to believe that geo-tagged tweets signal atypical activity, such as attending an event or travel, rather than the *home* location of its author. However, it is precisely this - the location where an individual lives, pays taxes and *votes* - which is of relevance to social scientists.

Second, to our knowledge, previous work has not provided a reproducible method along with free and open-source software shared with the research community at large. We see this as an important step toward improving reproducibility and transparency in social science research involving Twitter data. Furthermore, giving researchers access to geo-located Twitter data provides them with a tool for producing more robust, reliable and informed research in the future. We address these issues in this paper by synthesising the most predictive elements of previous geo-locating methods into a new, integrated, hybrid method of geo-locating any sample of Twitter users, and produce a ready-to-use software package for other researchers to apply in their own Twitter-based social science research. Furthermore, we investigate the usefulness of GPS-derived data for benchmarking geo-locating Twitter users.

In this paper we describe a hybrid hierarchical method for geo-locating Twitter users, which uses individual-level profile location metadata and parses it into five distinctive components of a *location estimate* - country, administrative area, municipality and latitude-longitude coordinates. If these data are not fully available or parseable for a given user, we extract a user's reciprocal friend-follower network (RFF), and estimate a user's location as a function of locations in this network, using modal hierarchical classification. We validate our approach in a two-step process using high-quality validation data obtained from a representative survey, as well as a large Twitter user sample with high-certainty user-location labels. Further, we describe a method for calculating classification certainty for any given location estimate.

Following this, we provide an empirical application of our method on a highly salient and contemporary political topic: the geo-spatial dynamics of Democratic presidential candidate mentions on Twitter leading up to and throughout the 2020 Democratic presidential primary season. Using a dataset of over 2 million unique users who tweeted about candidates from January 1st through May 31st 2019 as our starting point, we use location estimates obtained from our method to understand dynamics of candidate attention and support throughout the campaign. This analysis is further strengthened by adding more recent candidate mentions by the same users for the period between January 1st to the end of February 2020 to our analysis, allowing us to study how public opinion on Twitter develops at the user-level. In summary, we find that different candidates appeal to different archetypes of users, which, when condensed and aggregated, paint a broadly similar picture to measures of the geographic variation of election-relevant public opinion obtained through traditional means, such as polling and actual election results.

4.2 Literature Review

When aiming to geo-locate Twitter users, it is useful to clearly define the desired outcome variable. Geo-locating can refer to different levels: ***home location*** or ***tweet location***. The two differ, in that locateable tweets (either by means of geo-tagging, or by specific reference to a local point of interest) do not necessarily refer to a given user's home location, but rather to the location where a tweet was sent from, or a location it is referencing. The focus of this paper is on estimating users' *home locations*.

Three main approaches to inferring Twitter users' home locations have been discussed in the literature: using *tweet content*, *tweet metadata* and *user networks*. ***Tweet content*** approaches rely on the assumption that the text of users' published content contains signals which can be used to predict users' locations. Despite variations between applications, these methods typically train a classifier with user-level tweet text as input data to predict the user's location. Early studies used simple Machine Learning models constructed to predict users' self-reported locations as a function of tweet text (Cheng et al., 2010; Eisenstein et al., 2010; Hecht et al., 2011). These approaches yield maximum accuracy levels around the 50%-mark, with wide error margins (measured in distance from the "real" location - between 100 and 1000 miles). Later research introduced more complex models, such as using geo-tagged tweets to develop "language models" trained on content specific to geo-tagged tweets (Kinsella et al., 2011). Here, a list of possible locations, derived from geo-tagged tweets, is compiled with an associated probability of any given word which occurs in the corpus having "originated" from a given source location - i.e. the word's co-occurrence with a given location. This is then used to provide a probability of any tweet in a corpus being from a location in a corpus. Other authors attempted to incorporate the interactional nature of the platform into a tweet-content based location estimation algorithm which assumes dialogue-derived tweet text to be more likely to contain signals pertaining to users' home locations (Chandra et al., 2011) than tweets in general. The authors suggest that the accuracy of their approach exceeds the state of the art for content-based approaches by 10%, and "the accuracy for RBPDM [the most predictive model] was 58.88% with an error distance of 300 miles" (Chandra et al., 2011, p. 843). Finally, Ren et al. (2012) use "Inverse Location Frequency (ILF) and Remote Words (RW) filters - to identify local words in tweets content", and users' self-reported, parseable locations as ground truth. The method has a maximum accuracy of 56.60% within a 100 mile radius of true home locations at the municipality level.

Overall, the evidence suggests that content-based approaches may be able to serve a supplementary role when other methods are insufficient, but are not useful enough by themselves if one's goal is to geo-locate a large proportion of any given sample of Twitter users within reasonable error bounds. This is likely primarily due to the fact that users' text content is very noisy, and there is no way of knowing *a priori* which elements of a user's text content relate to geography. Furthermore, the problem of type mismatching home versus tweet location is pertinent when using content-based methods, regardless of which source of ground truth data is selected: there is no way of knowing if users tweet about real-world places when they are visiting an unusual place or if they are documenting their everyday life, suggesting that classifiers built on this framework are likely to contain unquantifiable levels of bias.

A different approach to geo-locating twitter users employs ***tweet metadata***. The key drawback of metadata-based approaches is the inconsistent availability and completeness of tweet- and user-level metadata, especially in regard to the prevalence of users' self-reported locations. Nonetheless, user-supplied locations are the simplest and most intuitively trustworthy data which can reveal users' locations. Mislove et al.'s (2011) seminal paper on identifying individual-level demographics for Twitter users introduced the concept of parsing self-reported location text through a geo-locating webservice such as Google Maps in order to systematically extract location-relevant information. Expanding on this, Han et al. (2013) use metadata-based methods, including self-reported location and users' time zones in conjunction with tweet content to geo-locate twitter users. However, adding all non-self-reported location measures into the estimation algorithm only improves accuracy from .40 to .49, while reducing median error distance from 92 to 9 miles (p. 9). Mahmud et al. (2014) use a content/metadata-based hierarchical approach of estimating users' home locations, resulting in a maximum accuracy of .58 at the city-level and .66 at the state level. This highlights the point that metadata-based location classifications can be improved by incorporating content-based classification into a hybrid methodology.

Based on these findings, metadata-based approaches give maximum accuracy rates around the 50%-65% - mark, depending on the location level which is being classified. However, this should be regarded with a certain degree of skepticism, given that both Han et al. (2013) and Mahmud et al. (2014) use locations extracted from geo-tagged tweets as ground truth data. However, the evidence suggests that location-specific profile metadata is the most reliable source of user-level location-specific signals, and is only then not useful when such a datum does not exist or is not germane for a given user.

Finally, several papers have employed information extracted from users' ***networks*** of social connections on Twitter as a means of geo-locating them (Backstrom et al., 2010; Davis Jr. Clodoveu A. et al., 2011; Sadilek et al., 2012; Jurgens, 2013; Compton et al., 2014; Rodrigues et al., 2016). At the core of network-based approaches to geo-locating social media users lies the assumption of *homophily* as a governing principle of humans' social ties (e.g. McPherson et al., 2001). If - as evidence suggests - this effect translates to interpersonal connections on Twitter, it is feasible to use information derived and inferred from users' networks to estimate socio-demographic attributes of said users, such as their home location. Different studies have used different conceptualisations of how best to estimate closeness or distance within a social network, and hence how to infer a target user's unobserved home location as a function of their neighbours'. Backstrom et al. (2010) used the population of U.S. Facebook users who provided home addresses in a network-based model of home location estimation for U.S. Facebook users. Using a maximum-likelihood model, the authors were able to accurately geo-locate 69.10% of users within a radius of 25 miles (Backstrom et al., 2010, p. 69) of their true home location. This study comes with the caveat that a) publicly available Twitter data are significantly less granular than *home addresses* and b) there may be top-level platform effects contributing to these findings[1] - such as Facebook users' readiness to share personal data. McGee et al. (2011) conducted a large-scale study of Twitter networks, investigating which network features (e.g. unidirectional following, reciprocal follower-friend connections, mentions) describing the relationship between users on Twitter best predict geo-spatial closeness/distance. The authors found that reciprocal friend-following relationships are the strongest predictors of closeness, and thus can be equated with the term 'friendship', whereby other forms of social ties, especially asymmetric connections such as mentions or unidirectional following tend to be predictors of distance. The authors suggest this describes the dual nature of Twitter as a platform: a social network and a news/content-delivery and consumption platform. Davis Jr. Clodoveu A. et al. (2011) employed a framework for estimating Twitter users' locations through network of reciprocal follower-friends, in line with McGee et al.'s (2011) findings. The authors use geo-tagged tweets from Twitter users who previously tweeted about Dengue Fever in Portuguese as ground truth, expanding their sample by recursively broadening users' networks starting with a small number of original users. Overall, precision for this approach reached a maximum of .40, while the best recall measure was 0.91. Jurgens (2013) leverages users' "ego-networks" to estimate locations, using a combined approach of label propagation and nearest neighbors in the ego-networks, as well as geometric medians of three-person reciprocal networks. Coordinate-based location data are assigned to 2,043,252 users[2] (p. 276), which are used as training data. While the authors state that only 0.70 % of tweets in their sample are geo-tagged, they are able to assign median locations of all geo-tagged tweets to 5.30% of users. The authors suggest that, using this machine learning approach, "nearly all of the users in the social network are located, with an estimated median error under 10km" (p. 281). Compton et al. (2014) use a similar approach, but expanding ground truth data to include self-reported locations and geotagged locations, while using a Total Variation Minimization algorithm to assign locations to unlabeled users. Rather than using RFF-networks, the authors use "bidirectional mention networks", which have the clear advantage that they can be distilled from an existing collection of tweets without having to access Twitter's API. Conversely, this also means that users not included in a seed collection cannot be assigned a location estimate using this approach, as all location estimates generated by the algorithm are derived from location data of the users in the set who *do* provide something, be it geo-tagged tweets or self-reported locations. Using leave-many-out validation, this approach labels 79% of users with geo-locations, with a median (mean) error distance of 6.38 km (289 km) (p. 399). It is important to note that this approach only works for users who @-mention other users in a seed collection, and not those users who only have unilateral (or no) mentions in their timeline. Rahimi et al. (2015) demonstrate that network-based methods generally outperform content-based methods: they require a smaller number of hyper-parameters in the authors' logistic regression model, use less memory and computing resources, and are not prone to the sensitivity of textual features and regularization settings. However, for users with few connections, a text-based approach achieved better accuracy. They found that a hybridization

[1] Furthermore, the researchers' role as Facebook employees allowed them to access data which would not be available for reproduction for the community at large - thus providing a useful showcase of the potential of leveraging networks to make individual-level demographic estimates, but not a replicable method

[2] the total number of users with geo-tagged tweets, extracted from a 10% sample of the full Twitter stream for 6 months in 2012

of the two components worked better than either method used independently, especially for those low-connected users. Rahimi et al. (2015) also proposed a label propagation approach based on Modified Absorption and similarly show that incorporating text-based priors and the removal of "celebrity" nodes (highly-mentioned Twitter users, who tend to have a disproportionately large number of connections) improved network-based results. Finally, Rodrigues et al. (2016) introduced a method combining network- and content data for location estimation. The authors achieve accuracy levels between 0.77 and 0.80 for a test set of users from 3 Brazilian cities, and 0.64 to 0.66 for a larger set of users from 10 Brazilian cities[3] (p. 34).

We have introduced the core approaches to estimating Twitter users' home locations: using tweet *content*, *metadata* or *networks*. On the whole, the evidence suggests that the different approaches vary in their classification accuracy, but it is hard to make generalizations given the different nature of datasets the methods were tested on. Overall, hybrid approaches have shown higher levels of accuracy than single-method approaches. Hence it is surprising that - to our knowledge - there has not been a study combining network- and metadata-based approaches. This paper introduces such a method. Furthermore, we argue that the reliance of the majority of papers in this field on locations derived from geo-tagged tweets as ground truth data may be problematic. While there has been little research into how users who geo-tag tweets differ from those who do not - it is such a rare phenomenon that assuming that such users are representative of Twitter users as a whole is likely to be false, and is likely to be biasing the results of such approaches.

Finally, it is important to note that these cited methods of geo-locating Twitter users provide explorations of their approaches on example datasets. We argue however, that for this practice to be useful for scientific research as a whole, it has to be easily transferred and applied to *any* sample of users. We fill this gap with our method, which requires no supplementary data to estimate home locations for any set of Twitter users.

[3] While the authors do not explicitly state this, their ground truth dataset seems to be identical to that used by Davis Jr. Clodoveu A. et al. (2011) - they use Brazilian geo-tagged Twitter users tweeting about "Dengue" in 2011.

4.3 Method

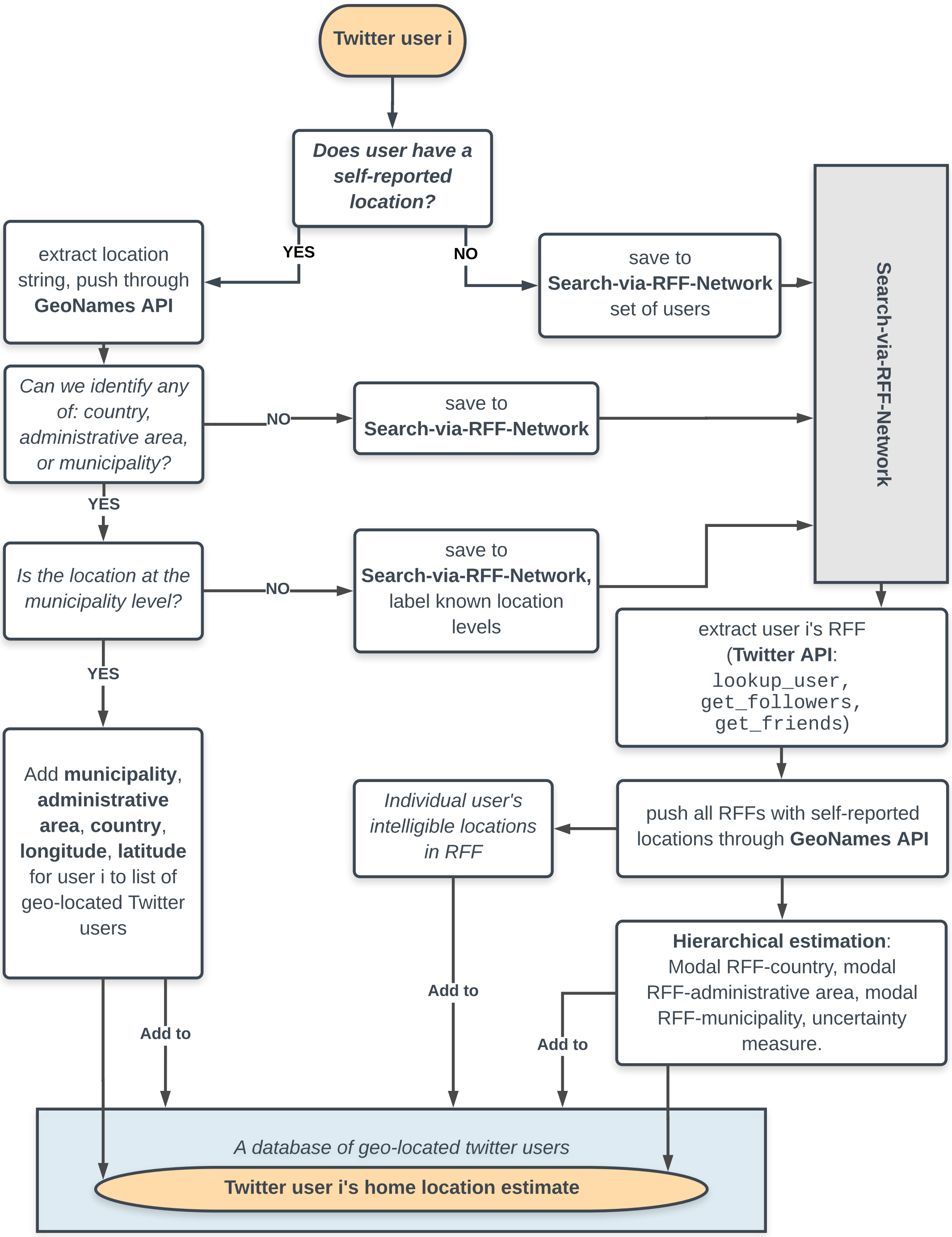

Twitter user i
Does user have a self-reported location?
YES
NO
extract location string, push through GeoNames API
save to Search-via-RFF-Network set of users
Search-via-RFF-Network
Can we identify any of: country, administrative area, or municipality?
NO
save to Search-via-RFF-Network
YES
Is the location at the municipality level?
NO
save to Search-via-RFF-Network, label known location levels
extract user i's RFF (Twitter API: lookup_user, get_followers, get_friends)
YES
Add municipality, administrative area, country, longitude, latitude for user i to list of geo-located Twitter users
Individual user's intelligible locations in RFF
push all RFFs with self-reported locations through GeoNames API
Hierarchical estimation: Modal RFF-country, modal RFF-administrative area, modal RFF-municipality, uncertainty measure.
Add to
Add to
Add to
A database of geo-located twitter users
Twitter user i's home location estimate

Figure 4.1: The 'Finding Friends' geo-locating pipeline

Our method for geo-locating Twitter users is a hybrid approach incorporating metadata and the reciprocal friend-follower network (RFF) of a given non-protected active twitter user. Our approach differs from existing ones in that it is designed for geo-locating any (random) sample of Twitter users, rather than providing a proof of concept on a purposefully selected sample.

Figure 4.1 schematically depicts the individual steps of our hybrid approach for geolocating any given Twitter user (referred to as *pipeline* below). Besides the assumption of homophily in reciprocal friend-follower Twitter networks, our pipeline presupposes that user-provided locations are the most reliable data on users' home locations accessible through the Twitter API - whenever they are available[4]. These core assumptions shape the data collection and analysis steps throughout our pipeline, e.g. determining which publicly available Application Programming Interface (API) endpoints are accessed to retrieve available data: Twitter's public Search API (https://developer.twitter.com/en/docs/api-reference-index), specifically the *get_user, get_friends, get_followers* endpoints, and the GeoNames Webservice for Geocoding and reverse geocoding (http://geonames.org). Furthermore, we store retrieved data, such as parsed locations in the form of profile-location-string key-value-pairs in a MySQL database, allowing for efficient retrieval of previously collected information for estimating new users' geo-locations. This way, costly API calls are reduced to a minimum.

The format of the pipeline's output is a) three hierarchical location string parameters - country, administrative area (e.g. a US state, a Canadian province, a Spanish autonomous region) and municipality (city, town, or village); *and*, where available, b) a pair of longitude-latitude coordinates which unambiguously define a geographic point on the earth. Previous research - typically using geo-tagged tweets as ground truth - has argued that longitude-latitude coordinates are the most useful output format (see e.g. Jurgens et al., 2015), as it allows for the calculation and estimation of a classified location's likely error radius and allows for a clearer operationalisation of "distance". However, if a user is only locateable at the country or administrative area level, the use of coordinates implies an unwarranted degree of certainty, which is likely to introduce bias when using such data to estimate other users' locations. Hence, our main focus is on name-based identifiers of geographical locations at three levels.

Census-list matching users with self-reported profile locations

Step one of the pipeline is to determine if a user's profile metadata features any text in the self-reported location field, and furthermore, if this text provides parseable location information down to the municipality level. As Twitter does not require this free-form text field to be completed by users and does not enforce any rules as to what can be entered, a considerable number of users have uninformative location information, such as 'The Universe', 'your mama's house' or '#Trumpland'. Furthermore, even if a given self-reported location string does contain valid location information, it may not cover all three possible location levels, (e.g. 'California' or 'Scotland'). We determine how informative self-reported location strings are by first matching these strings with a database of place names - location-string key-value pairs, collected using the GeoNames API. If this fails, we use the GeoNames API to parse the user's location string.

If the returned location is at the municipality-level, its location is returned in the format depicted in Table 1. It includes the users' unique Twitter user-id ('ID'), their profile location string ('String'), the country, administrative area and municipality returned from passing the user's location string through our place name database and the GeoNames web service

[4] While this is certainly open to debate, we suggest that the voluntary nature of supplying such data, and its permanence as a profile feature rather than a tweet-level feature, make it the most intuitively trustworthy heuristic for classifying user-level home locations.

('Country', 'Admin' and 'Mun.', as well as longitude and latitude coordinates ('Long', 'Lat'). However, if the output obtained from parsing the user's location metadata is either empty, uninformative or incomplete, the user is added to a list of users whose location is to be estimated using the network-component of our pipeline. Furthermore, if a user's self-reported location provides information only at the country- or admin-level, the network-approach location estimation is conducted contingent on this information in order to estimate a municipality for this user.

ID	**String**	**Country**	**Admin**	**Mun.**	**Long**	**Lat**
123456	'nyc'	United States	New York	New York	-74.006	40.714

Table 4.1: Output format for parsed locations (Census-list match)

The Census-List component of our geo-locating pipeline typically assigns approximately 33% of users within a given sample of Twitter users with a location at the municipality-level[5].

Estimating users' home locations with reciprocal follower-friend networks (RFF)

For the approximately two thirds of Twitter users for whom a complete home location cannot be obtained using the census-list method, we leverage the vast amount of data available in users' networks. Specifically, we extract information from the user's RFF-network. As stated above, a user whose profile metadata yields uninformative location data is added to a queue of users flagged for location-estimation using the network-component of our pipeline. For a user $\boldsymbol{u}$ in this queue, we collect user-ids for up to 5000 users whom $\boldsymbol{u}$ follows, as well up to 5000 users who follow $\boldsymbol{u}$. We then take the intersection of these 2 sets, i.e. $\boldsymbol{u}$'s observed RFF. Sets of followers and friends, can vary widely in size, and can be exhaustively large, which is why our algorithm cuts off data collection at 5000 friends and 5000 followers.[6] For every user in $\boldsymbol{u}$'s observed RFF, we then repeat the procedure outlined above - we collect user metadata for each member of the observed RFF, extract RFF-users' self-reported location strings and use our database of location-string/parsed-location key-value pairs and the GeoNames API to attempt to identify the subject's location. Table 2 shows example output for an RFF-location database. In this case, user $\boldsymbol{u}$ has an RFF of size N=8, of which two users' (3, 6) metadata did not return usable location strings[7]. Table 2 also highlights the flexibility of using the GeoNames API for parsing strings to informative locations: strings as vague as 'nyc' or as specific as 'union sq' are classified correctly as New York, New York, while strings with errors/typos, such as 'Lodnon' are correctly parsed as London.

There are multiple ways of calculating $\boldsymbol{u}$'s estimated home location given the RFF-location database, such as modal location estimation, hierarchical modal location estimation

[5] This figure is derived from running numerous samples of Twitter users through our pipeline which are not explicitly discussed in this paper. While these samples do not form a representative picture of Twitter users in general, we are still confident that these proportions hold true across Twitter as a whole, as the figures are consistent across a diverse range of differently selected samples.

[6] This is in line with Twitter's API limits, which does not allow collection of more than 5000 user-ids per call to the API endpoint. We refer to this observed RFF below as user $\boldsymbol{u}$'s RFF, but in fact it may be a subset of the user's RFF.

[7] This may be due to: no location string, incomplete location string, or protected accounts (these allow access to metadata only to users they have granted access to)

ID	String	Country	Admin	Mun.	Long	Lat
1	nyc	United States	New York	New York	-74.006	40.714
2	nyc	United States	New York	New York	-74.006	40.714
3	NA	NA	NA	NA	NA	NA
4	Chicago, IL	United States	Illinois	Chicago	-87.623	41.881
5	Brooklyn, NY	United States	New York	New York	-74.006	40.714
6	NA	NA	NA	NA	NA	NA
7	union sq	United States	New York	New York	-74.006	40.714
8	lodnon	UK	England	London	-0.118	51.509

Table 4.2: An example RFF-location database entry

or weighted hierarchical modal location estimation. Different approaches may perform better given different underlying network configurations - consider, for instance, a user with an RFF-network of size 1000, with 100 edges located within California, but into many different **municipalities**, and several smaller regional clusters around the world. Here, a clustering approach will likely place a user in California, while a modal approach may place the user in the most common location in the set, regardless of whether any regional clustering is observed around that location. However, this is an empirical and indeed answerable question, which, after testing multiple configurations on randomly sampled Twitter users, has led us to conclude that hierarchical modal estimation is the approach which performs best on average.

For the example depicted in Table 2, this means that the modal location at every location level is determined, and this is then contingent on the next-lower level. In the example case in Table 2, the modal country is 'United States' (5/8). Given this, the 8^{th} location in this observed RFF is no longer incorporated into the estimation for user ***u***'s administrative area, as it does not lie within the modal country. In this example, New York is the estimated administrative area (4/5), and New York is the estimated municipality(4/4).

Hence, for this case, our pipeline will output and store *u*'s location as 'New York, New York, United States, -74.006, 40.714'. The advantage of storing this location estimate is that it can later be utilized when aiming to estimate user ***p***'s location, if ***u*** is in ***p***'s RFF.

4.4 Validation

How well does this geo-location method perform? Before putting it to work, in this section we first: a) evaluate the accuracy of the method, b) explore the conditions under which it works best, and c) use the latter to measure the uncertainty of our geo-location estimates. We focus on evaluating model performance for Twitter users in the United States, our primary substantive area of interest. However, the same validation strategy can easily be generalised to other contexts.

4.4.1 Model Accuracy

We use a two-fold strategy to validate our method. First, we evaluate the accuracy of the census-list component of our pipeline by comparing the actual and estimated locations for a set of 1,091 Twitter users who provide location information in their Twitter profile and for whom we actually know their country (United States) and administrative area-level (state) location, as they participated in a representative survey of US adults conducted by YouGov for NYU's SMaPP Lab in 2016 .[8] We produce country and administrative area-level estimates using the census-list component for 847 of the 1,091 users (82%). The accuracy of these country and admin-level estimates is very high: 92% and 83%, respectively.

Then, we use a different strategy to evaluate the performance of the network-estimation component of our pipeline. Since this is the most novel part of the method, we want to have detailed information about its performance and test its accuracy on a much larger sample. We use the census-list component of our pipeline to locate tens of thousands of random Twitter users[9], and then we draw a random sample of 30,000 users for which the census-list matching component returned location estimates in the United States. We take advantage of the high accuracy of the census-list estimates and treat these 30,000 country and admin-level census predictions as the gold standard for evaluating the network method. In this application, we employ *weighted hierarchical modal location estimation*. We recommend this approach especially for samples of users from one country, as is the case in the below example. We weight distributions of administrative area-level locations in a target user's RFF using the real-world (or real-Twitter[10]) population distribution at that level. In practice, for the US case, this requires importing population distribution figures at the state-level, either derived from the census, or from another source (in this case, we use the state-level distribution of random US Twitter users located using the census-component of our pipeline), and subtracting the observed proportion in a given RFF from the expected proportion. For instance, if a user's RFF contains 15% of users from California, and 3% of users from Wyoming, the non-weighted approach would place this user in California, as the algorithm is agnostic to underlying population sizes of administrative areas. However, the empirical proportion of Twitter users from Wyoming is below 1%, while California's is also 15%. Hence, using weights means we place this user in Wyoming. We further perform log-transformation on the post-weight proportions.

[8] These 1,091 Twitter users were selected from a panel survey conducted before the 2016 U.S. election. Among other questions, 3,000 nationally representative U.S. respondents were asked for their: a) Twitter handle and b) U.S. state of residence. 1,693 provided a handle and 1,091 of these turned out to have a self-reported location string linked to their Twitter profile.

[9] The random Twitter users were selected by generating random numeric user IDs and then checking whether the users existed.

[10] I.e., the empirical geographic distribution of the population of US Twitter users

State	Prop. U.S.	Prop.	Predicted	Precision	Recall	Top 3 Mistaken States
California	0.115	0.127	3166	0.55	0.61	NY (n=178), TX (n=137), FL (n=129)
Texas	0.105	0.110	2744	0.68	0.71	CA (n=119), FL (n=79), PA (n=39)
New York	0.069	0.048	1191	0.76	0.53	CA (n=44), FL (n=30), NJ (n=29)
Florida	0.068	0.062	1534	0.64	0.58	CA (n=71), NY (n=55), TX (n=45)
Georgia	0.042	0.042	1044	0.62	0.62	FL (n=53), CA (n=46), NY (n=45)
Illinois	0.039	0.030	752	0.81	0.63	CA (n=26), TX (n=13), CO (n=13)
Ohio	0.035	0.029	731	0.80	0.68	NY (n=18), CA (n=17), FL (n=12)
Pennsylvania	0.034	0.028	709	0.80	0.66	NY (n=23), NJ (n=12), CA (n=12)
Michigan	0.030	0.026	641	0.81	0.70	CA (n=20), FL (n=13), IL (n=10)
Massachusetts	0.026	0.022	549	0.79	0.67	CA (n=26), NY (n=17), TX (n=9)
Tennessee	0.025	0.025	626	0.68	0.67	FL (n=23), CA (n=22), TX (n=20)
North Carolina	0.024	0.021	525	0.74	0.67	GA (n=18), FL (n=18), CA (n=17)
New Jersey	0.024	0.019	469	0.76	0.59	NY (n=28), CA (n=12), PA (n=11)
Arizona	0.020	0.018	440	0.64	0.57	CA (n=43), TX (n=18), FL (n=15)
Alabama	0.018	0.018	460	0.72	0.73	GA (n=33), CA (n=12), TX (n=10)
Indiana	0.017	0.016	395	0.72	0.66	GA (n=13), IL (n=11), TX (n=10)
Kentucky	0.016	0.015	381	0.77	0.76	OH (n=10), CA (n=10), NY (n=7)
Minnesota	0.016	0.014	350	0.79	0.69	CA (n=11), NY (n=7), FL (n=7)
South Carolina	0.016	0.016	411	0.67	0.70	GA (n=17), CA (n=15), FL (n=13)
Maryland	0.016	0.015	372	0.68	0.63	CA (n=13), DC (n=11), NY (n=10)
Virginia	0.016	0.014	361	0.61	0.56	WV (n=44), NY (n=12), NC (n=10)
Colorado	0.016	0.014	358	0.60	0.54	CA (n=23), NY (n=17), TX (n=14)
Louisiana	0.015	0.018	443	0.65	0.74	TX (n=26), CA (n=24), GA (n=15)
Missouri	0.015	0.012	306	0.70	0.57	TX (n=11), CA (n=11), IL (n=9)
Washington	0.015	0.014	353	0.61	0.58	CA (n=28), DC (n=19), TX (n=16)
Washington, D.C.	0.014	0.024	607	0.27	0.48	CA (n=80), NY (n=49), TX (n=38)
Oklahoma	0.013	0.013	320	0.76	0.75	TX (n=20), CA (n=11), GA (n=5)
Nevada	0.013	0.015	377	0.49	0.58	CA (n=47), TX (n=20), FL (n=17)
Wisconsin	0.012	0.013	321	0.62	0.66	IL (n=17), CA (n=15), MN (n=13)
Oregon	0.012	0.012	290	0.63	0.60	CA (n=19), WA (n=13), TX (n=12)
Iowa	0.011	0.011	286	0.71	0.73	CA (n=15), IL (n=10), MN (n=7)
West Virginia	0.011	0.009	236	0.39	0.35	OH (n=14), TX (n=13), NY (n=13)
Mississippi	0.010	0.013	313	0.58	0.76	TX (n=19), TN (n=15), FL (n=14)
Arkansas	0.010	0.011	278	0.61	0.66	TX (n=21), GA (n=9), TN (n=6)
Connecticut	0.009	0.008	203	0.59	0.55	NY (n=21), CA (n=10), MA (n=7)
Utah	0.008	0.009	222	0.52	0.62	CA (n=18), NY (n=11), FL (n=10)
Kansas	0.007	0.008	211	0.54	0.69	MO (n=29), TX (n=13), CA (n=9)
Nebraska	0.006	0.007	174	0.71	0.78	CA (n=8), TX (n=5), FL (n=5)
Hawaii	0.004	0.006	145	0.36	0.57	CA (n=22), NH (n=10), NY (n=9)
Maine	0.004	0.006	146	0.26	0.40	CA (n=18), TX (n=11), NY (n=11)
New Mexico	0.003	0.005	115	0.36	0.49	CA (n=19), TX (n=8), NY (n=7)
Rhode Island	0.003	0.007	162	0.28	0.70	CA (n=21), NY (n=15), MA (n=11)
Montana	0.003	0.005	128	0.30	0.52	CA (n=15), TX (n=11), NY (n=10)
Idaho	0.003	0.007	186	0.26	0.66	CA (n=22), TX (n=13), GA (n=11)
New Hampshire	0.003	0.007	164	0.27	0.54	TX (n=15), NY (n=14), CA (n=14)
North Dakota	0.002	0.004	98	0.28	0.69	CA (n=18), FL (n=8), MN (n=5)
Vermont	0.002	0.004	94	0.27	0.64	CA (n=12), NY (n=10), TX (n=4)
South Dakota	0.002	0.006	141	0.25	0.67	CA (n=15), TX (n=13), NY (n=9)
Alaska	0.002	0.005	117	0.20	0.38	CA (n=16), NY (n=12), FL (n=11)
Delaware	0.001	0.003	78	0.22	0.50	CA (n=7), FL (n=6), TX (n=4)
Wyoming	0.001	0.008	191	0.07	0.50	CA (n=31), TX (n=20), FL (n=14)

Table 4.3: Accuracy when classifying the state location of users known to be in the U.S.

We proceed to using the network-estimation component to estimate a country and admin-level location for these 30,000 users. First, we collect these users' RFF networks. The average user has a network with 693 friend-followers, whereas the median user has an RFF-network of size 287. About 4,000 users do not have *any* reciprocal friend-followers, and so we can only generate a network-based prediction for 25,778 of the users, approximately 86% of them. The precision of country-level estimates is 97%, meaning that the United States was the modal network country location for 24,914 of the 25,778 users. The precision of the state estimates is 62.3%. In Table 3, we provide detailed information about the accuracy of the administrative area-level network estimations. Both the precision and recall is higher than 50% for 35 of the 51 administrative units, higher than 60% for 22 of them, and higher than 70% for 5 of them: Alabama, Kentucky, Oklahoma, Iowa, and Nebraska. Two main error types stand out. First, we misplace some people in their neighbouring state. For example, the two top states in which we incorrectly place users from Virginia are West Virginia and Washington D.C. (see the bottom rows for the sixth column from the right). Moreover, the most common error involves misplacing people into the most populous states, California, New York, Texas, and Florida. However, this type of error would have been even more common if we had not weighted the state location of the users' RFFs.

Benchmarking model performance versus geotag-derived location classification

As mentioned before, the majority of the existing literature on geo-locating Twitter users employs data derived from users' geo-tagged tweets. This requires users to activate the GPS (global positioning system) of the device from which they are tweeting, and actively opt in to having their tweet tagged with the latitude-longitude coordinates of their current location. Then, the tweet will contain parsed geographical information on the web (e.g. "Manhattan, NY"), as well as granular latitude-longitude coordinates when retrieved from the Twitter API.

We are sceptical of the utility of such data for the purpose of classifying Twitter users' home locations, for two reasons. Firstly, geo-tagged tweets are an exceedingly rare phenomenon (with approximately 0.85% of overall tweets geo-tagged (Sloan et al., 2013)), suggesting that geo-tagged tweets either signify an atypical event or an atypical user. In other words, our prior is that users avail themselves of this feature when they want to signal an unusual activity, such as participating in a protest or attending a concert, and not to share exact coordinates to their home. For this purpose, we suggest that using geo-tagged tweets as baseline data (e.g. for training algorithms) for geo-locating Twitter users may be biasing any resulting classifications, and thus may be less useful than user-level profile location data.

However, this is, at least in part, an empirical question. Hence, we compare the accuracy and performance of location estimates obtained from geo-tagged tweets for the same sample of 30,000 US Twitter users. We collected users' most recent tweets (the most recent 3200 tweets per user, collected in April 2020), retaining only tweets with geo-tags. In total, there were ***5564*** users, or 18.6%, with at least *one* latitude-longitude coordinate-pair in our sample. The mean number of coordinate-pairs per user (out of users with at least one geo-tagged tweet) was 108.6, while the median number was 20. Figure 4.2 depicts the distribution of the percentage of geo-tagged tweets out of each users' timeline, for the 5564 users in our sample who had at least *one* geo-tagged tweet. Even in this specific subset of users, the vast majority of users chooses to geo-tag only a small subset of all their tweets - 71%, or n=3985 of users geo-tag fewer than 5% of their tweets, while only a total of 2 users enabled geo-tagging for *all* their tweets. This further supports our notion of geo-tagging as a behaviour for atypical tweets by atypical users.

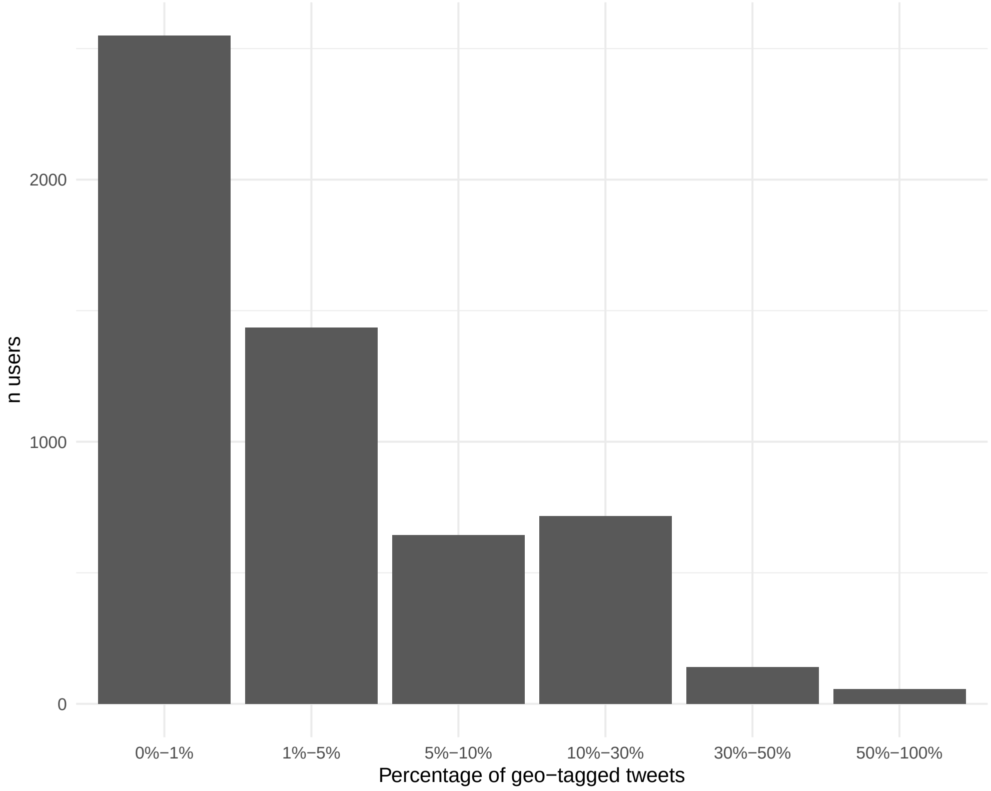


Figure 4.2: Users with geo-tagged tweets, grouped by the percentage of their tweets that are geo-tagged

A coordinate-pair does not automatically yield a real-world place. Further, there is a question of how to aggregate multiple coordinate-pairs into one 'average' location. For this purpose, we pursue two distinct strategies of parsing users' coordinate-pairs into human-readable locations. First, we use the haversine distance formula to calculate a centroid-point between *n* coordinate-pairs, as well as the area, in square kilometres, covered by all the users' coordinate-pairs. Further, we also calculate the median coordinate-pair (i.e. median latitude, median longitude) for each user. If a user has 2 or fewer coordinate-pairs, we choose the *first* coordinate-pair. We then parse all centroid-coordinates and median coordinates through the GeoNames reverse geocoding API, with a maximum error radius of 100km. This accounts for error and noise caused by aggregating multiple coordinate-pairs into one. Then, in order to assess the accuracy of this method, we compare parsed locations from these methods with our census-list derived gold standard set of user-level locations.

The centroid-method classifies an administrative-area location for 4937 users, or 16.5% of the sample, while the median-method classifies administrative-area locations for 4505 users, or 15% of the sample. Table 4 shows location-level accuracy numbers for both methods, broken down by sample and sub-sample proportions. While the centroid method outperforms the median method at the country-level, at least numerically, the median method has a clear edge at the administrative area and municipality level. This is likely explainable by

	Centroid			Median		
	n	*% classified*	*% of sample*	*n*	*% classified*	*% of sample*
≥ 1 coord-pair	5564		18.54	5564		18.54
parsed locations	4937	88.73	16.45	4505	80.96	15.01
correct: Country	4556	92.33	15.20	4312	95.71	14.39
correct: Admin	2585	52.39	8.62	3264	72.45	10.89
correct: Mun.	168	3.40	0.56	384	8.52	1.28

Table 4.4: Accuracy breakdown for geotag-derived location classification; number/percentage of correctly classified users

the vast area covered by many users who geotag their tweets, meaning that the centroid coordinate-pair may be somewhere in the sea, while the median pair will be close to the cluster with *most* coordinate-pairs. However, it is also apparent that either method of deriving home location estimates for Twitter users from tweet geotags is not fit for purpose. At the administrative area-level, a maximum of 11% of our 30,000 user sample gets correctly classified.

This clearly demonstrates that the geo-tag method is *not* suitable as training/source data for geo-location algorithms which otherwise avail themselves of other data - If centroid-locations are chosen (as is the case in e.g. Compton et al. (2014) or Jurgens et al. (2015)), only half of the set of users whose location is even classifiable will be correctly classified. A better, yet nonetheless insufficient approach would be the median method. We argue that this finding brings into question reported accuracy and performance figures for existing geo-locating methods for Twitter users.

4.4.2 Model Uncertainty

Under which conditions does the hierarchical network model perform best? We believe that network features can help us predict accuracy. For example, we expect larger location modes to be positively correlated with accuracy. So, to be more likely to correctly place a user u in Virginia with a network $g_u = \{\text{VA,VA,VA,VA,PA,PA,IN}\}$ than a user p with a network $g_p = \{\text{VA,VA,PA,PA,OH,IN,WA}\}$. As a first step towards building this uncertainty measure, we collect the network features noted in Table 5 for all users in the validation set.

Feature	**Description**	**Example**
Network Size	The number of RFFs in a user's network	7
Mode Size	The number of RFFs in the most frequent location	4
Second Option Size	The number of RFFs in the second most frequent location	2
Number Other Options	The number of non-modal location in which RFFs live	2
Number Other People	The number of RFFs in non-modal location	3
Non-Modal Dispersion	A measure indicating the dispersion of RFFs' locations*	.5

**Note*: Formula to calculate Non-Modal dispersion: $\frac{\sum_{j=1}^{Z}(|x_j-(N/Z)|)}{Z}$, where N is the *Number Other People*, Z is the *Number of Other Options*, and x_j is the number of RFFs in each 'other option' j.

Table 4.5: Network features used to predict the accuracy of the network predictions.

We then construct a statistical model to predict accuracy and thus measure the uncertainty of location estimates. Since we are agnostic towards the "correct" specification and the functional form of this statistical model, we proceed by: a) fitting a large number of models (n = 1,974 logistic regressions)[11], b) evaluating the performance of these models, and c) estimating a final ensemble model that predicts accuracy as a function of the accuracy predictions of the best models. As a result, we obtain an ensemble model that we can use to measure uncertainty by estimating the probability of any network-estimated location to be correct.

Figure 4.3 summarises the accuracy (precision and recall) of the 1,974 statistical models we fit. Larger dots indicate that a greater number of models have that particular precision and recall. We observe a strong trade-off between the two measures: models with high precision tend to have low recall, and vice versa. We also observe models with state-level fixed effects to be more accurate. The red triangle indicates the precision (85%) and recall (85 %) of the final ensemble model, which takes into account the predictions of the ten most precise models. The high performance of this final ensemble model means that we are able to produce reliable uncertainty estimates for our geo-location predictions. Further information about the relationship between individual network features and model uncertainty can be found in appendices.

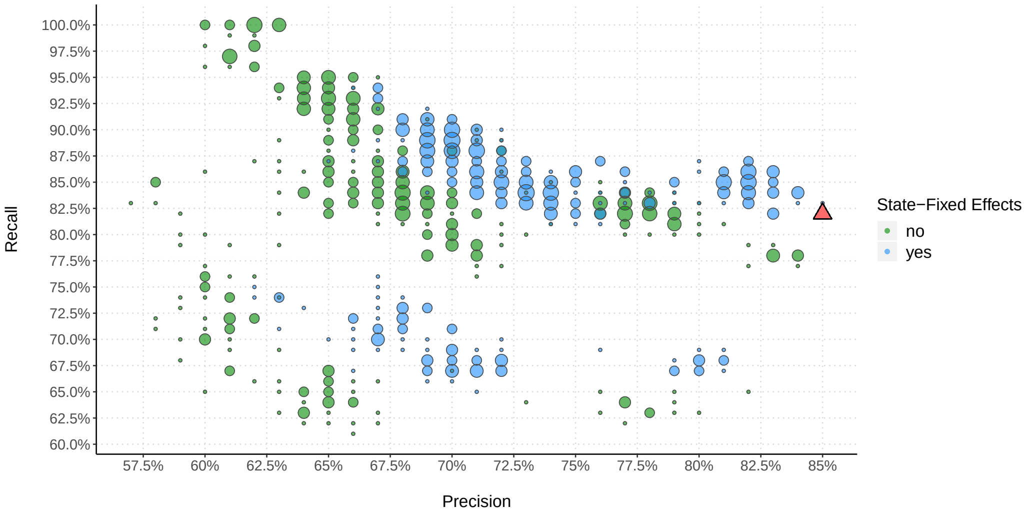


Figure 4.3: Precision and Recall for 1,974 models predicting the accuracy of the network-estimated locations as a function of network features.

4.4.3 Prediction uncertainty

The ultimate goal of geo-locating a given set of Twitter users is to perform some type of analysis on them for which user-level location is relevant. Hence, we want to make sure that we are confident about the location of such users. We use the ensemble model we trained

[11] For each of the 6 network features, we take into consideration their linear but also their potential logarithmic and quadratic effect on the outcome, increasing the number of input model features from 6 to 18. Then we take into consideration all possible 1-to-3 variable combinations for a total of: $\sum_{n=1}^{3} \frac{18!}{n!(18-n)!} = 987$ models. Finally, we run each model twice, with and without state-fixed effects, increasing the final number of models to estimate to 1,974.

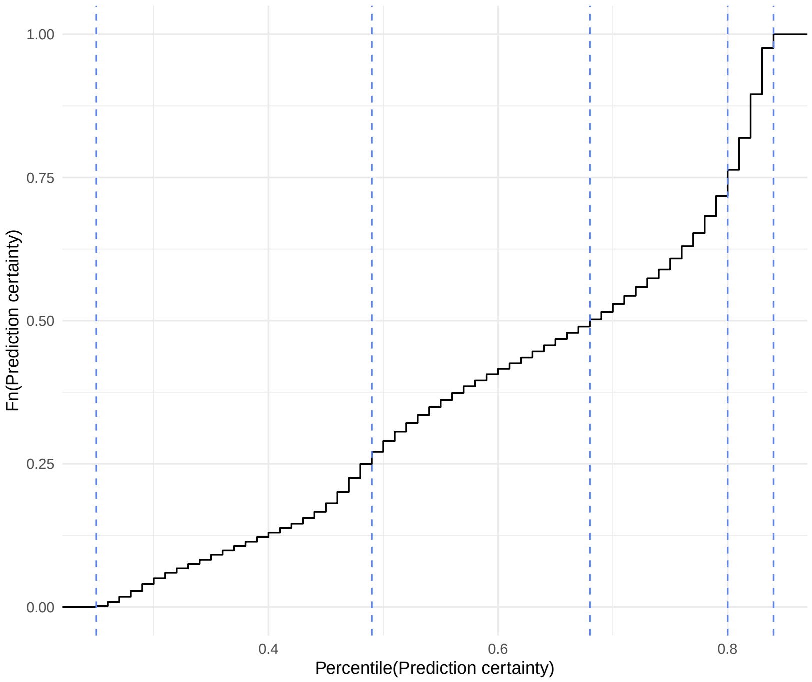


Figure 4.4: ECDF (Empirical Cumulative Distribution Function) showing the distribution of the probability of each of our 25,778 network-based state predictions to be correct (Quantiles in blue)

to predict the probability of our 25,778 admin-level predictions to be correct. As shown in Figure 4.4, we observe that approximately 60% of predicted certainty estimates are over the threshold of .6, while half the estimated scores are at least greater than .68. However, we can also see that 25% of predicted scores range below the certainty value of .5. The mean of all certainty estimates for our network-validation sample is 0.64, the median is 0.68 and the standard deviation is 0.17.

These probabilities not only allow us to measure the uncertainty of our predictions, they also allow us to decide how confident we want to be about the location of the users we use for analysis. Inevitably there is a trade-off between the number of located users and how confident we are about their location. Researchers can decide to be more or less ‘cautious’ depending on their needs.

4.5 Empirical Application: Exploring the Geo-Spatial Dynamics of Tweeting about the 2020 Democratic Primaries

We now give a comprehensive and relevant example of how our method can be applied in a political science research scenario. We study users tweeting about the 2020 US presidential election, focusing specifically on tweets mentioning Democratic party presidential candidates. Our focus lies in understanding geographic patterns of online public opinion and political participation as they relate to elections, and further, to investigate how these geographic patterns vary depending on individual user-level characteristics as well as the temporal dynamics of a long, geographically disaggregated campaign.

4.5.1 Data

We collected a data set of **7,117,065** tweets mentioning any declared or assumed Democratic party candidate for US president in the time period between January 1st and May 31st 2019. The candidates included in this tweet collection are:

> *Amy Klobuchar, Andrew Yang, Bernie Sanders, Beto O'Rourke, Bill DeBlasio, Cory Booker, Elizabeth Warren, Eric Garcetti, Eric Holder, Eric Swalwell, Jay Inslee, Jeff Merkley, Joe Biden, John Delaney, John Hickenlooper, Julian Castro, Kamala Harris, Kirsten Gillibrand, Mitch Landrieu, Pete Buttigieg, Seth Moulton, Steve Bullock, Tim Ryan, Tom Steyer, Tulsi Gabbard.*[12]

With the benefit of hindsight, having witnessed a number of candidates drop out of the race, and others on this list never actually declaring their candidacy, while others' bids proved largely insignificant, we reduced the number of candidates in our analysis to eight, namely:

> *Amy Klobuchar, Andrew Yang, Bernie Sanders, Elizabeth Warren, Joe Biden, Kamala Harris, Pete Buttigieg, Tom Steyer.*

For this set of candidates, we collated the set of unique candidate-mentioning tweets into a set of unique users with attached mention-counts for each relevant candidate. Having dropped users who only mentioned excluded candidates, we were left with 1,766,739 unique users who tweeted *at least once* mentioning one of the eight candidates in the time period of interest. We geo-located these users using our hybrid pipeline. This set includes users from **255 countries and territories**. Of these users, ***1,265,202 users, or 71.6% are from the United States***. The next most common countries are the UK (2.7%), Canada (2.2%), while users from France and Venezuela comprise approximately 0.8% of the sample. We suggest that the large proportion of non-US users likely stems from a number of sources:

- It highlights the global relevance of US politics. While only US citizens are allowed to participate in the US presidential election, people around the world follow it. Hence, this is the most plausible source of non-US tweeting.

[12] A tweet was treated as mentioning a candidate if it contained any of the collection terms listed in any of the first name / last name combinations of the candidates; as well as *only* last names for the following candidates: Warren, Biden, Sanders, Buttigieg, Klobuchar, Yang. While a tweet mentioning a candidate *may* contain the candidates' Twitter handle, we did *not* include their handle specifically in this analysis, but rather focused exclusively on their names.

- There is a significant potential that a proportion of tweets within this sample stem from automated accounts (bots), or trolls and spammers. While there exist some approaches to identifying such accounts, this paper does not focus on that.

- A small number of users are likely to have been mis-classified at the country level, especially if they used uninformative language for their self-reported profile location. However, as documented above, this number is exceedingly low, and thus not relevant to this large-n analysis.

As our source data for mentions of Democratic presidential candidates is from 2019, we felt this analysis would benefit from also seeing where users continued to tweet about a Democratic presidential candidate in **2020**. This allows us to investigate the geo-spatial patterns of mentioning political candidates *over time*. For instance, we may be interested to see where users who tweeted about a certain candidate who later dropped out of the race a) shift towards tweeting about a different candidate or b) stop tweeting.

For this purpose, we applied the same filters to pertinent Twitter collections from 2020 (January 1st 2020 - February 12th 2020), and further subset the resulting mention-collection to only contain mentions by users already in the original 2019 collection. This left a dataset of **707,495** unique users with at least *one* mention of one of the pertinent candidates, within the 2020 time frame. Out of these users, **543,561** are from the US (located at least to the state-level). It is striking that fewer than half of the users who tweeted about a presidential candidate in our 2019 set were still mentioning one of our eight Democratic candidates in our 2020 set.

In order to further understand the composition of our sample, we calculated location certainty estimates for all user-level location classifications using the model trained to validate the network-component of our pipeline. Table 6 shows the state-level location classification certainty by certainty thresholds. Notably, this sample has considerably higher overall location classification certainty than our validation sample. The mean certainty value for US users in this sample is 0.94 (median: 1, standard deviation 0.16)[13]. As prediction certainty is estimated using the features of a given users' network, this suggests that the key features associated with a reliably classified location are more frequently present in this 1.2m US users sample than the 30k random users sample. As the users in *this* sample were collected as a result of posting or sharing content on Twitter as opposed to randomly, this is an encouraging finding for prospective research applications of this geo-locating pipeline, as real-world social science research using Twitter data rarely employs randomly sampled Twitter data.

[13] These summary statistics were computed by dropping any missing ('NA') values; if NAs are treated as a location estimate with certainty==0, the summary statistics are as follows: mean: 0.76, median: 0.99, standard deviation: 0.39.

	State	$\geq$ **.5**	$\geq$ **0.6**	$\geq$ **0.7**	$\geq$ **0.8**	$\geq$ **0.9**	$\geq$ **0.95**
1	AL	92.64	91.08	90.09	88.12	74.80	67.15
2	AK	88.24	87.40	86.26	84.62	75.83	72.90
3	AZ	90.62	89.82	88.06	86.10	77.24	70.53
4	AR	90.24	89.41	87.66	85.64	72.93	66.85
5	CA	95.21	95.21	95.21	92.53	68.67	51.67
6	CO	94.32	94.15	92.14	91.14	86.14	82.67
7	CT	92.30	90.23	89.60	87.21	76.21	69.07
8	DE	84.49	83.05	80.23	75.40	55.32	50.08
9	FL	96.27	96.27	95.85	91.07	75.08	62.02
10	GA	93.48	92.58	92.05	89.60	76.50	68.06
11	HI	92.08	90.83	89.88	88.12	77.40	73.62
12	ID	90.45	89.12	87.76	85.68	76.57	73.34
13	IL	97.45	97.45	94.59	93.66	87.51	80.38
14	IN	93.56	92.50	91.00	89.19	80.02	72.13
15	IA	91.74	91.11	89.32	88.00	77.18	70.99
16	KS	90.95	89.45	88.47	85.98	68.90	61.04
17	KY	93.22	92.89	90.99	89.46	79.66	72.01
18	LA	92.67	91.98	90.82	89.62	75.47	67.67
19	ME	90.84	89.53	88.46	86.58	76.97	73.11
20	MD	92.89	92.28	90.45	88.87	74.62	64.22
21	MA	96.50	93.60	92.90	91.56	83.15	74.56
22	MI	96.06	93.14	92.54	90.58	81.06	72.57
23	MN	91.95	91.31	89.89	87.85	79.73	71.84
24	MS	92.80	91.47	90.57	88.56	71.73	62.83
25	MO	93.87	93.79	91.33	90.66	82.64	75.38
26	MT	91.38	90.75	89.51	87.39	78.73	75.93
27	NE	92.62	91.38	90.08	88.88	78.82	72.97
28	NV	93.94	92.98	92.10	91.11	83.55	78.19
29	NH	91.86	91.47	90.50	88.81	82.05	78.66
30	NJ	96.38	94.17	94.02	91.93	85.65	77.91
31	NM	93.43	93.17	92.25	91.35	85.88	82.84
32	NY	97.04	97.03	97.03	93.96	85.92	74.67
33	NC	94.50	94.37	93.09	92.22	83.04	75.12
34	ND	89.89	88.87	87.98	86.89	80.46	77.05
35	OH	97.06	95.25	94.32	93.63	87.07	80.06
36	OK	95.08	93.70	93.54	92.24	86.39	80.88
37	OR	94.17	92.41	91.68	89.70	81.12	75.10
38	PA	97.16	94.74	94.12	93.20	87.32	80.17
39	RI	87.53	86.96	86.09	84.55	77.29	72.37
40	SC	95.54	94.29	94.02	92.81	84.65	78.98
41	SD	88.02	86.90	85.19	81.71	70.15	65.31
42	TN	93.95	93.47	92.64	90.64	79.61	72.22
43	TX	96.41	96.41	96.41	94.76	69.12	59.60
44	UT	93.51	93.39	91.64	90.30	84.68	79.75
45	VT	91.13	89.38	87.97	86.17	75.89	71.94
46	VA	93.99	92.35	92.01	90.19	76.38	64.51
47	WA	95.07	94.54	92.83	91.37	84.72	78.63
48	DC	91.31	89.69	88.74	86.97	72.67	69.05
49	WV	95.29	95.10	94.18	92.83	88.00	85.62
50	WI	92.31	91.86	90.10	88.64	77.67	70.72
51	WY	92.75	91.83	90.92	89.69	85.19	83.66

Table 4.6: Percentage of location classifications by certainty thresholds, 1.26m US users

4.5.2 Analysis

Location and mention distribution

While non-US users in our sample are clearly an important subset in their own right, our focus is on US users. For this purpose, it is useful to contextualise *where* US Twitter users in our sample live. Our naive prior is that the state-level distribution of this sample will differ from a random sample of US Twitter users, and more so from the census-derived, *actual* population distribution, as, a) people who tweet about politics are not 'typical' Twitter users (see e.g. Bekafigo and McBride, 2013), and b), people who tweet about the Democratic party presidential campaign are a more left-leaning subset thereof. Geography is heavily correlated with politically salient individual-level attributes such as ideology. Therefore, we expect so-called 'blue states', i.e. states that lean Democratic, to be over-represented in our data, while Republican-leaning 'red states' should be under-represented.

Figure 4.5 depicts over and under-representation by state of the 1,265,202 US users in our sample, versus a) the empirical state-level population distribution as measured by the 2017 US census, and b) the expected state-level distribution of a random sample of US Twitter users). It is apparent that both the real-world US population distribution *and* the Twitter distribution differ considerably from our sample. This is, on the whole, more pronounced for the census-derived distribution than the Twitter-derived one. Furthermore, the blue/red state hypothesis mostly holds true: California and New York are widely over-represented. However, there appears to be more of a population size effect than a red/blue effect, as Texas is also significantly over-represented, although it is still widely considered a red state. An outlier is West Virginia, a state with a low population which is considerably over-represented. As outlined in section 4, it is likely that some of this over-representation stems from misclassifications of users from Virginia, a significantly more populous, and 'bluer' state, into its neighbour. Overall, we can conclude that we are working with a highly specific sample of Twitter users, which differs greatly from both the general and the Twitter population.

Besides the geographic distribution of our sample, we are also interested in which candidate gets mentioned where, and how frequently. While the next section goes more in-depth into user-level tweeting/candidate-mentioning behaviour, it is useful to have an overview of state-level candidate mentions. For this purpose, Figures 4.6 and 4.7 outline the total number of mentions, across the entire time frame of the sample, from all users located to the United States. The figures are separated into four candidates each, for the sake of readability. Figure 4.6 features state-level mention totals for Joe Biden, Kamala Harris, Bernie Sanders and Elizabeth Warren, while Figure 4.7 features Pete Buttigieg, Amy Klobuchar, Tom Steyer and Andrew Yang.

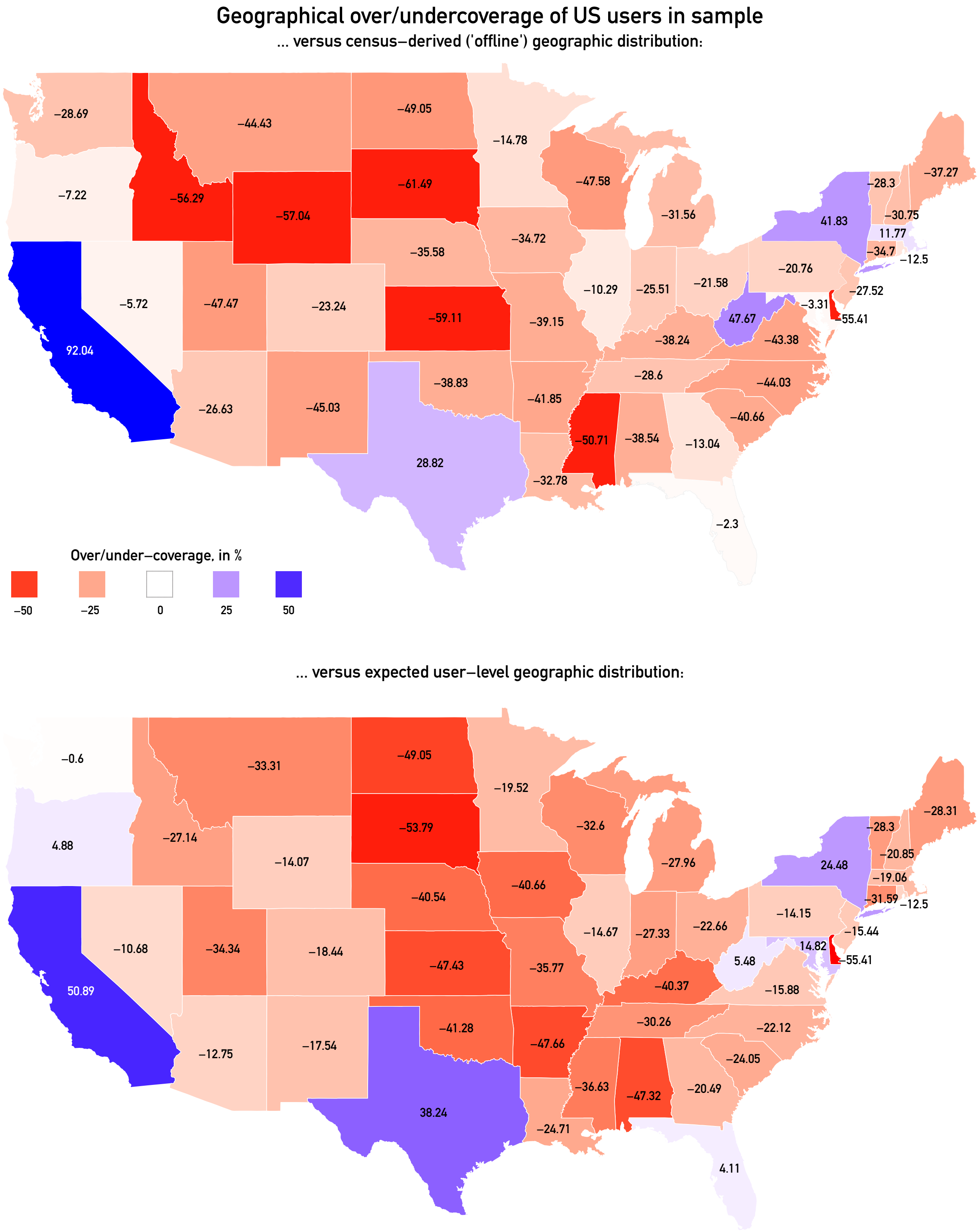


Figure 4.5: Over and undercoverage of sample compared to Twitter average and 2017 census.

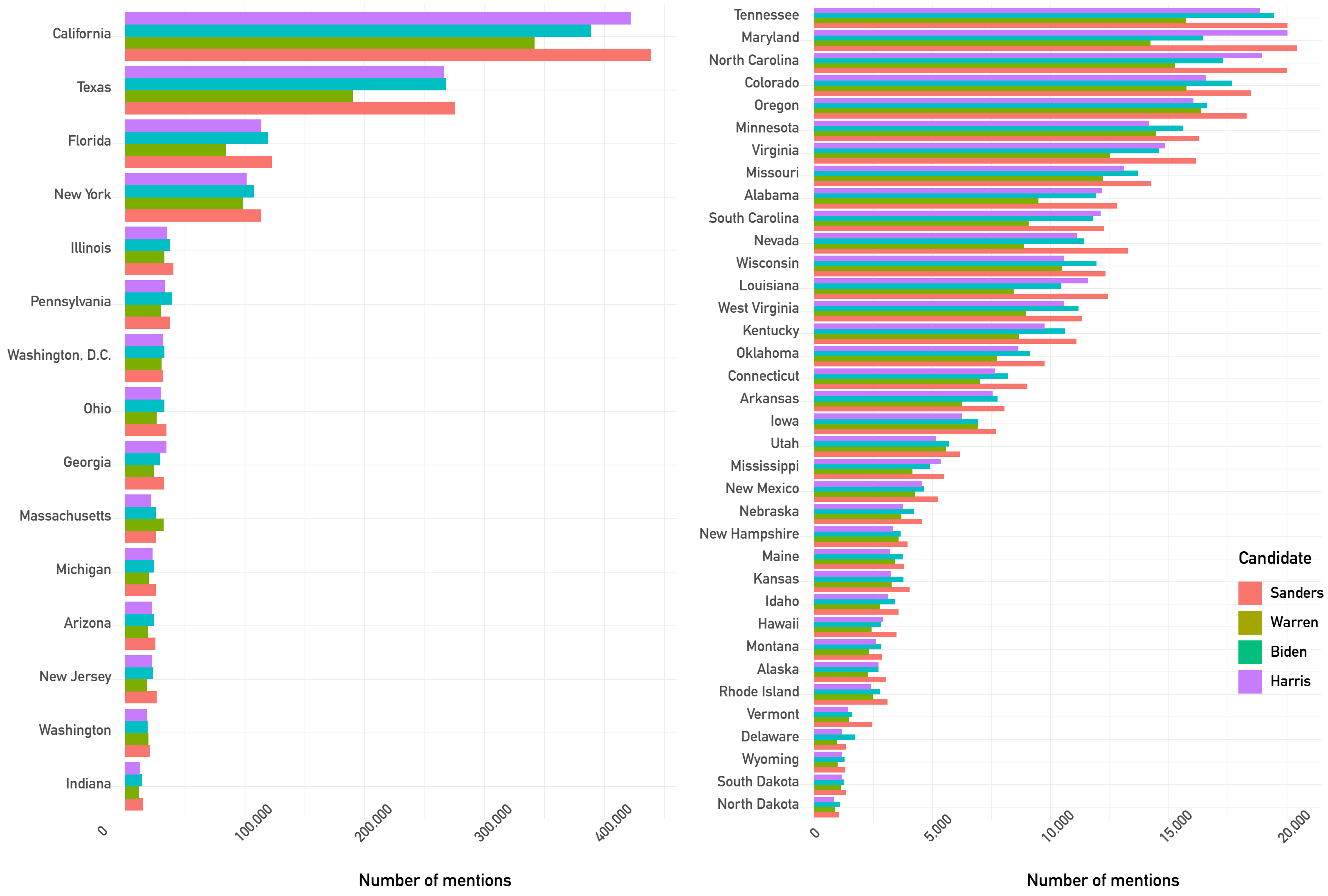


Figure 4.6: State-level mention distribution for Biden, Harris, Sanders, Warren

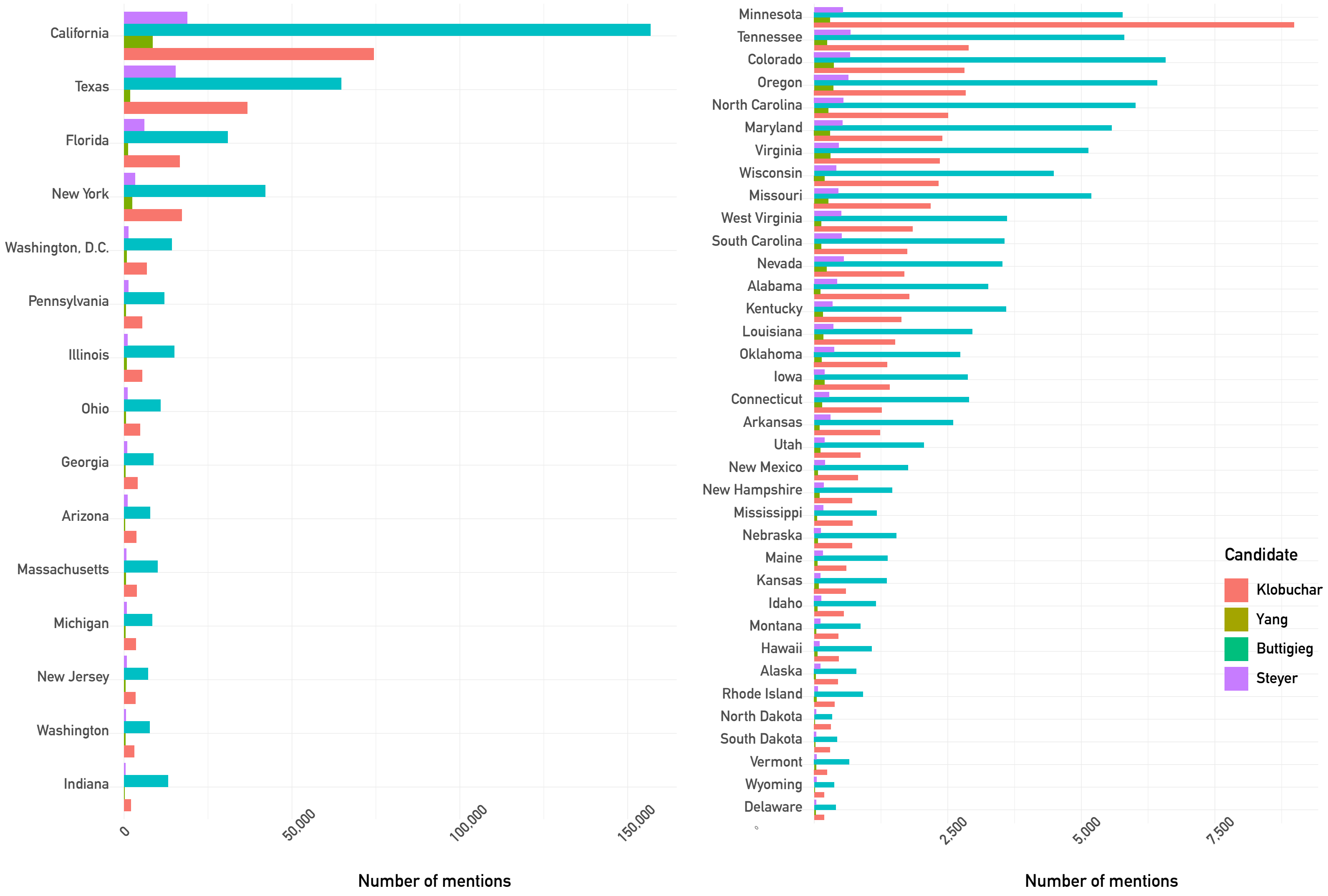


Figure 4.7: State-level mention distribution for Buttigieg, Klobuchar, Steyer, Yang

It is apparent that the candidates in Figure 4.6 received considerably more mentions in our sample than those in Figure 4.7. For instance, Bernie Sanders has more mentions in North Carolina, his state with the 20th-highest number of mentions, than Tom Steyer in California - his state with the highest number of mentions. Furthermore, it stands out that Bernie Sanders is ahead in terms of raw mention numbers in the vast majority of states. In order to contextualise this, and other candidates', mention totals by state, it is useful to also consider candidates' *total* mention numbers. This is depicted in Table 7.

	Candidate	***n* mentions**	**%**
1	Bernie Sanders	2070566	24.98
2	Joe Biden	1877920	22.65
3	Kamala Harris	1795194	21.66
4	Elizabeth Warren	1518067	18.31
5	Pete Buttigieg	624940	7.54
6	Amy Klobuchar	292816	3.53
7	Tom Steyer	78122	0.94
8	Andrew Yang	31901	0.38

Table 4.7: Mention totals for selected candidates

Clearly, Bernie Sanders is ahead, in terms of the total number of mentions in tweets. This is the case both at the aggregate level, and at the state level for the majority of individual states. However, this is a misleading metric. Any Twitter user can tweet about Bernie Sanders - or any other candidate - as often as they wish, and Bernie Sanders, whose national profile is higher than that of most of his opponents, is known to have a strident following of devoted supporters on Twitter - all of whose conversations about their hero will have been captured in these data. For this purpose, we argue that a much more useful analysis of the mentions of presidential candidates in tweets is conducted at the user-level, which gives significantly less weight to prolific tweeters. We describe this analysis in the following section.

User-level analysis

So far, we have only looked at *who* users were tweeting about, and *where* they were tweeting from at the aggregate level. However, this leaves many of the individual-level user-characteristics of the 1.2 million users in our sample under-explored. Specifically, we want to learn more about what kind of users tweet about which candidates, and where. So, while the previous section was all about the candidates, this section is all about the users.

To begin, we divide the US users in our samples into certain informative archetypes of candidate-mentioning behaviour. To achieve this, we choose to segment the users by overall candidate mention frequency. The underlying assumption for this choice is that more (politically) engaged users - i.e. those users who mention presidential candidates a lot - are different from those who casually re-tweet one link about Joe Biden. Our goal is to learn more about these user archetypes, and how they vary in how they express political participation and public opinion on Twitter.

Figure 4.8 depicts the percentage of mentions a candidate receives from specific subgroups of users, sorted by mention frequency of any of the candidates over the course of the entire collection period. This acts as a proxy for political attention/engagement of users:

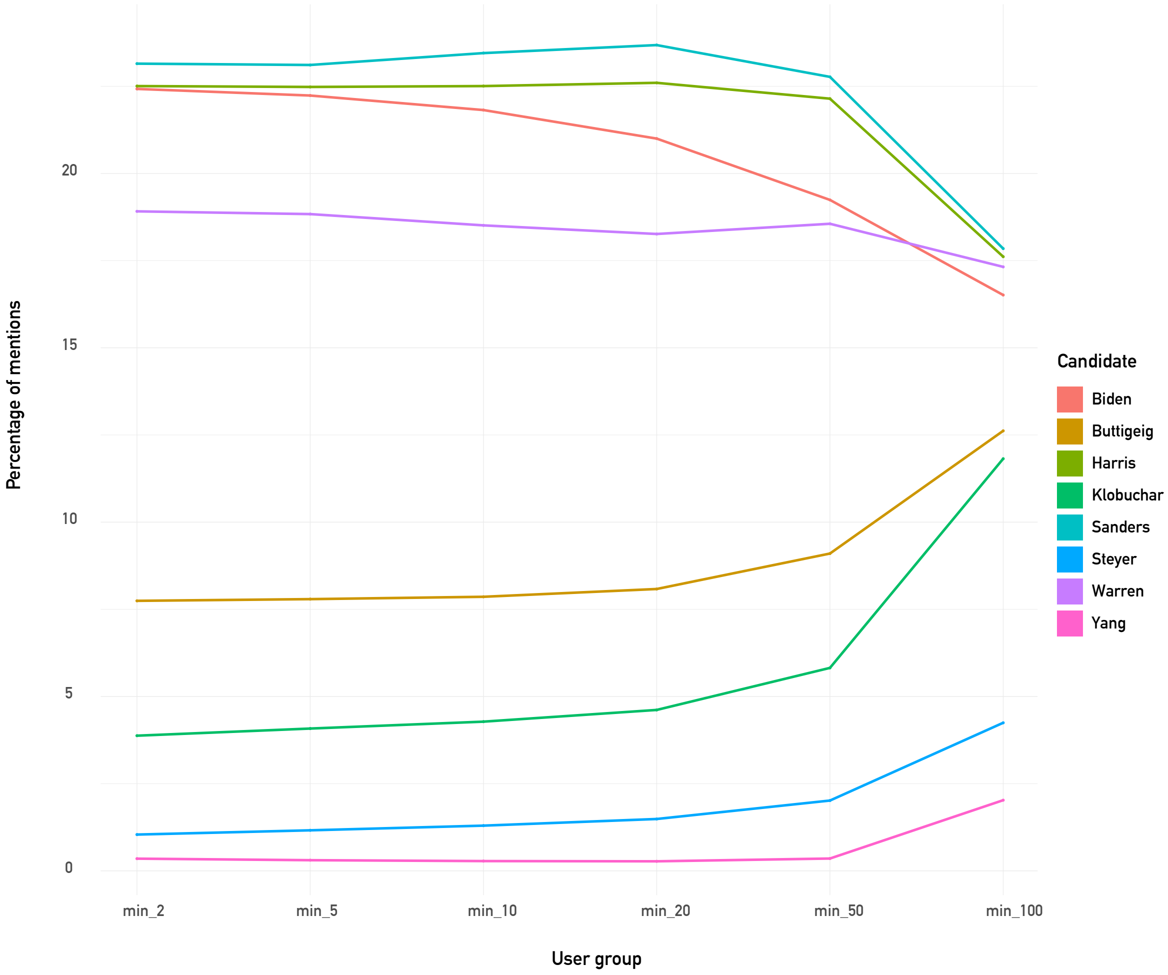


Figure 4.8: Candidate mention percentages by user archetypes, derived from overall user-mention frequency (user groups defined cumulatively)

users who only mention a candidate once throughout the 6-month period likely differ in how closely they pay attention to politics from users who mention candidates at least 50, or 100 times in the same period. It is important to analyse how mention totals shift for candidates when subset by user-archetype, as it may be an indicator as to how well candidates are connecting with specific groups of US twitter users.

The graph shows a clear pattern - there are 4 candidates - Bernie Sanders, Kamala Harris, Joe Biden and Elizabeth Warren - who have strong appeal with less engaged users, but their mention proportions decrease for more engaged users, with a particular dip for those users who mentioned candidates at least 100 times in sample 1. Conversely, the remaining candidates - Pete Buttigieg, Amy Klobuchar, Andrew Yang and Tom Steyer - have lower mention proportions for less-engaged users (ranging between under 1% for Yang and under 8% for Buttigieg), but significantly higher mention proportions for more engaged users. From this, we can deduce that a large proportion of Twitter mentions for candidates with mainstream appeal (here: Sanders, Biden, Harris, Warren) received a significant amount of their Twitter attention from less politically engaged users, whereas outsider candidates (here: Buttigieg, Klobuchar, Yang, Steyer) got most of their attention from politically engaged users. Furthermore, we can conjecture that users in our sample who tweeted a lot have higher political knowledge and attention than those who tweeted rarely, as they are aware of lesser-known candidates, and contribute a considerable share of their overall mentions.

Given that this paper is about the geography of Twitter users, and how this can better inform our understanding of public opinion as expressed through Twitter, this analysis results in the question of where different archetypes of users are tweeting from?

Having established the difference in mention proportions for different user groups in the sample, as defined by their frequency of tweeting about a candidate, it is further relevant to analyse the candidate preferences of these users. For this purpose, we computed the top candidate for every user who tweeted at least 5 times in the whole sample. We define 'top candidate' as a candidate that accounts for at least 50% of all of a user's candidate-mentions. Figure 4.9 depicts the candidate who accounts for the plurality of 'top candidate' status for users with at least 5 mentions in the entire dataset for a given state. Strikingly, this map paints a significantly different picture than Figures 4.6 and 4.7[14].

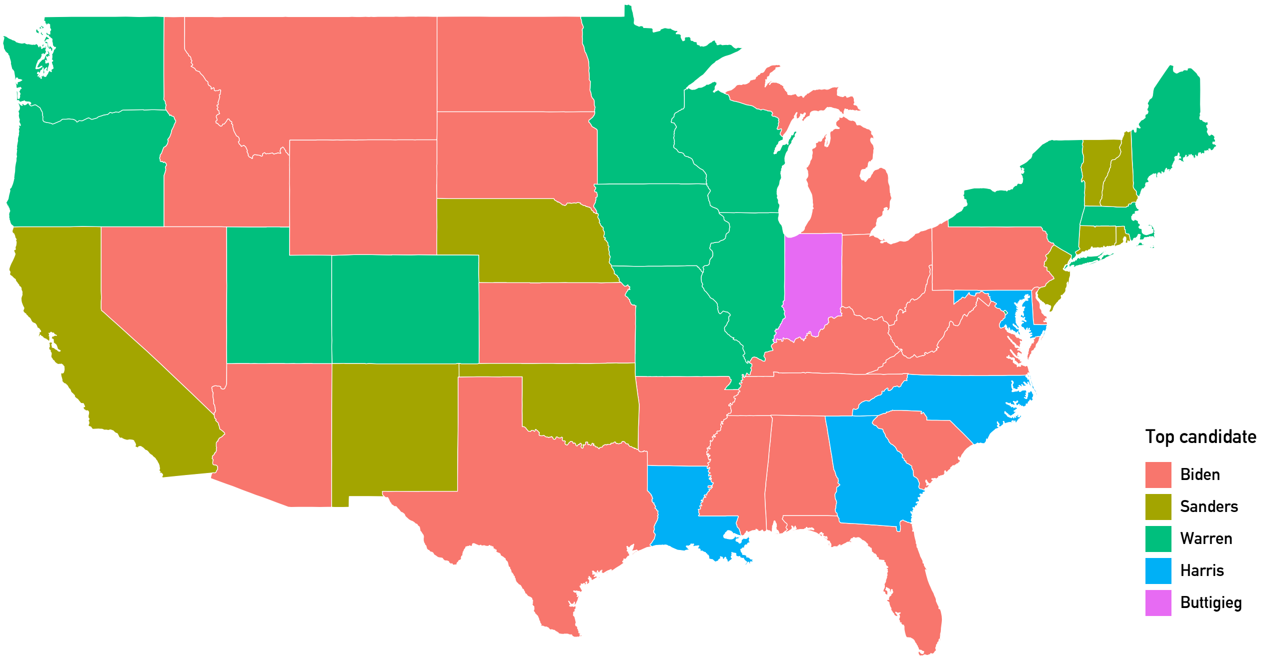


Figure 4.9: US Map depicting most common 'top candidate' by state, Data from Jan 1st 2019 - May 31st 2019

Here, Joe Biden is clearly the 'top candidate' in most states, and other candidates are ahead in certain states too, notably Pete Buttigieg in his home state of Indiana. Furthermore, Elizabeth Warren, whose overall mention numbers trail those of Sanders, Biden and Harris, is ahead in this top candidate metric for a significant number of states, many of which are traditionally considered as textbook 'blue states', such as New York, Washington, Oregon, Illinois or her home state of Massachusetts. A further interesting pattern is that of Kamala Harris being the top candidate in states with large African-American populations, namely Maryland, North Carolina, Georgia and Louisiana. This is consistent with what we know

[14] See Appendix A for a table with the full state-level percentages for each candidate

about the voting behaviour of black Americans, and their documented track record of near-unitary support of black democrats (e.g. Fairdosi and Rogowski, 2015) - at least if they are Democrats. While one may argue that these states also feature large white, rural and suburban populations, the fact that these groups are unable to overturn Harris' top candidate status in these states likely reflects the demographics of Twitter, which skew toward young people who live in urban centres - which also happens to be where most black people in Harris' states live.

To sum up, Figure 9 highlights the difference that a shift of the unit of analysis for analysing Twitter-based public opinion can make. While the tweet/mention-level analysis depicted in Figures 4.6 and 4.7 would suggest that Bernie Sanders is leading the field, the user-level analysis (albeit for users with at least 5 tweets in the whole collection) would indicate otherwise, suggesting that the population of moderately politically engaged to very engaged Twitter users are split between Biden, Sanders, Warren and Harris. Furthermore, there is a clear indication that the geographic variation in public opinion which is well-studied and documented across the political science literature, and reproduced in public opinion poll after poll is *also* reproduced in geographically dis-aggregated individual-level data gathered from Twitter. This is a highly encouraging finding for the field of Twitter-based public opinion research as a whole, which has been subject to considerable amounts of criticism due to its perceived inability to reliably reproduce established patterns of offline public opinion.

Where do users change their minds? Comparing data from 2020 to data from 2019.

As mentioned earlier, we have expanded the data for the same users for January 1st - Feb 12th 2020. This allows us to trace the dynamics of attention toward presidential candidates on Twitter, and examine the state-level variance for these processes at the user-level. Figure 4.10 shows a Sankey diagram depicting the flows of user-level top-candidate statuses by candidate, from the first period of mention data (1st January - 31st May 2019) to the second (1st January - 12th February 2020). It is immediately apparent that Bernie Sanders retains the majority of users who have him as his top candidate, specifically benefiting from people who tweeted mostly about Kamala Harris in period 1, and re-aligned their mentioning behaviour toward other candidates once Harris announced the termination of her campaign. While Biden also receives a sizeable proportion of previous Harris-mentioners, he also makes a net loss to Bernie Sanders, and his overall number of top-candidate-users is lower in period 2 than it is in period 1.

For the purpose of mapping the geographic variance of user-level change in attention to presidential candidates on Twitter, we are particularly interested in three distinct types of users. Firstly, we want to know where users who had Kamala Harris as their top candidate in 2019 switched the candidate they were tweeting about most, and to whom. Secondly, we are interested in those users who did *not* have a top candidate in 2019 but have one in 2020, where this change occurs, and to whom. Finally, we are also interested in the overall state-level distribution of top candidate statuses for the 2020 dataset (i.e. an updated rendering of Figure 4.9).

Figure 4.11 shows the most frequent 'top candidate' per state, for users who had Kamala Harris as their top candidate in the 2019 dataset. While most users in the majority of states moved their Twitter mentioning activity toward Joe Biden, Bernie Sanders also gains

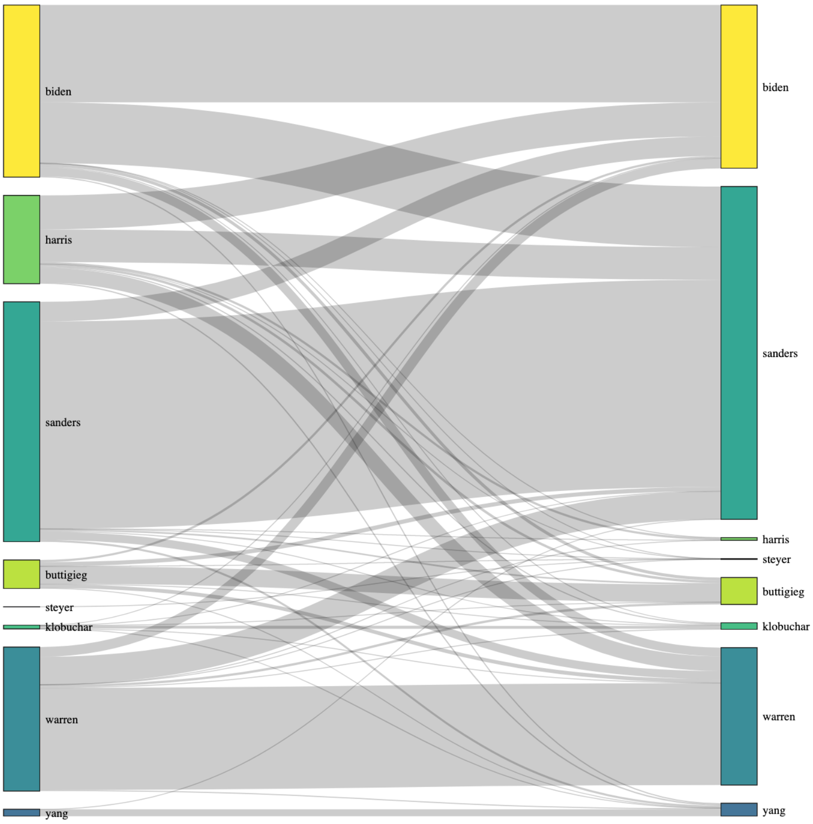


Figure 4.10: “Top candidate" flows for users with a top-candidate (min 50% of mention-tweets in set) in 2019 and 2020 tweet collections

mentions from Harris supporters in high-population states such as California and New York. Intuitively, it appears as though former Harris mentioners in more liberal states (California, Washington, Illinois, New York, Massachussets) shifted their mentioning activity toward Sanders, while those Harris mentioners in more conservative states (Texas, Florida, Georgia, the Carolinas) appear to shift to the relatively more conservative Biden. It is, of course, important to note that we do not claim to be measuring candidate approval, but rather a measure of regional *attention* received by a candidate - however, it is clear that candidate approval, or voting intention, or something similar to this is correlated with this measure of attention. It is encouraging that we can trace the flow of attention previously directed to Harris towards a more liberal candidates in liberal states, and to a more conservative candidate in more conservative states.

Figure 4.12 shows two top candidate maps. First, the most frequent top candidate by state, for users whose mentioning activity did not lead them to have a top candidate in 2019. There is a stark south-east/rest divide between Biden and Sanders. Biden achieves most frequent top candidate status only in a small number of southern states and Texas (conservative states), while Sanders leads throughout the rest. The second map is an updated version of Figure 4.9, i.e. containing *all users'* top candidate in the 2020 set, regardless of their status in 2019. It is striking how similar the two maps in Figure 4.12 are - besides Oklahoma.

For the users in our samples, it is apparent that we can trace a consolidation in attention towards presidential candidates, from a crowded field in 2019 to a two-horse race in early

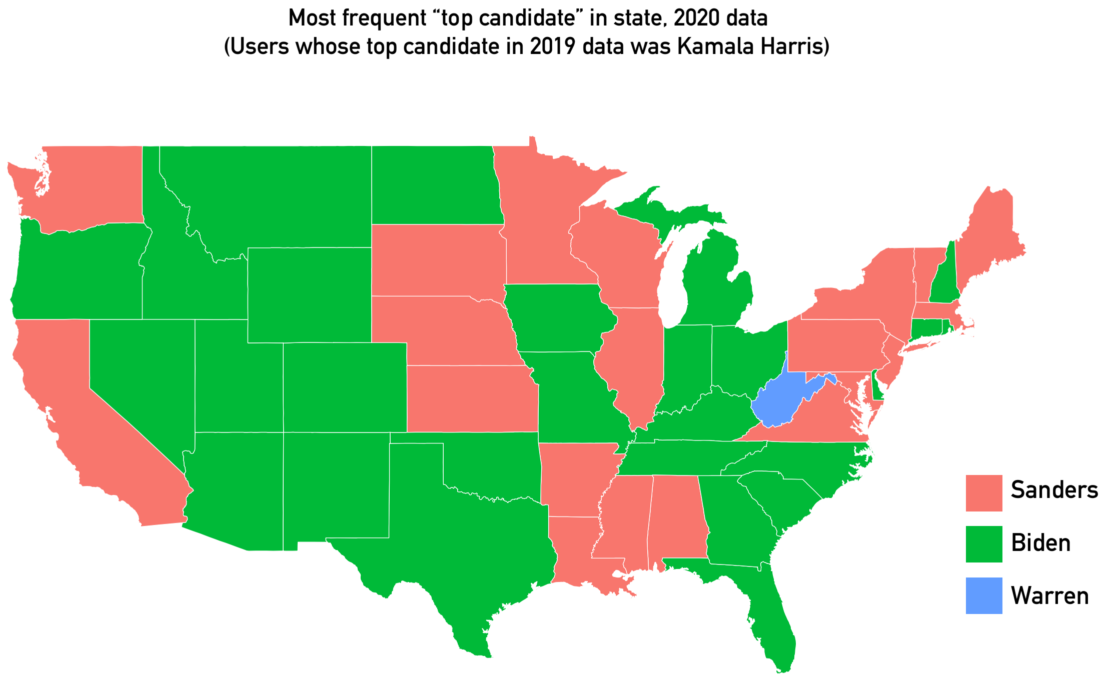

Most frequent “top candidate” in state, 2020 data
(Users whose top candidate in 2019 data was Kamala Harris)
Sanders
Biden
Warren

Figure 4.11: Most frequent user-level "top candidate" by state, 2020 set, for users whose 2019 top candidate was Kamala Harris (min 50% of mention-tweets in set)

2020. While it is important to note that the top candidate distribution would look very different had the source data been *anyone* mentioning a Democratic presidential candidate in the 2020 period, rather than users who mentioned candidates in 2019 ***and*** 2020, it is highly informative to have evidence that indicates a high degree of geographical and temporal volatility of public opinion on Twitter.

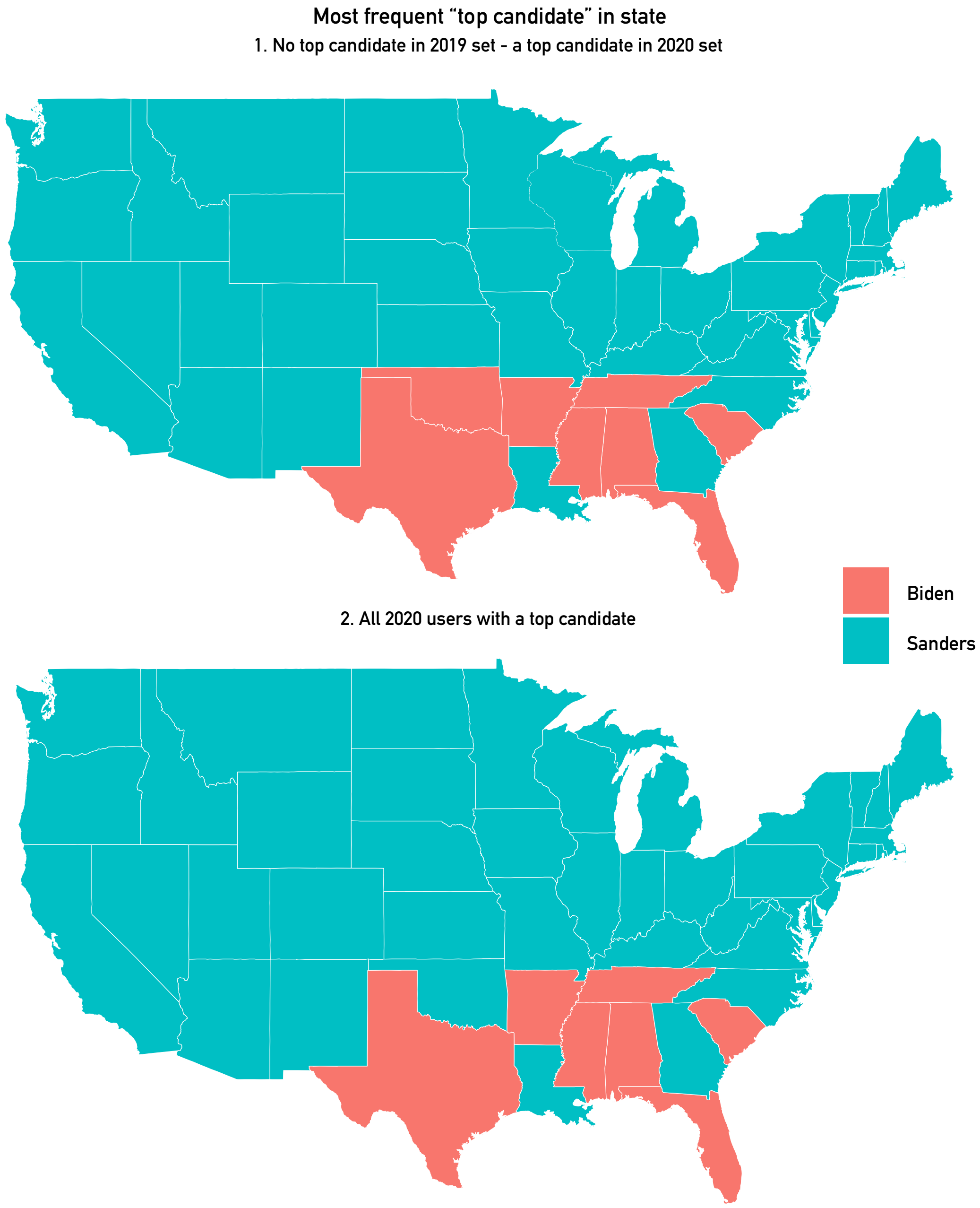

Most frequent “top candidate” in state
1. No top candidate in 2019 set - a top candidate in 2020 set
2. All 2020 users with a top candidate
Biden
Sanders

Figure 4.12: Most frequent user-level “top candidate" by state, 2020 set, 1 - for users who had no top candidate in the 2019 set; 2 - for all 2020 users with a top candidate.

4.6 Discussion

In this paper, we have outlined and demonstrated a fully reproducible method for attaching a user-level location estimate to any sample of Twitter users. This method leverages both user-level profile metadata (in the form of users' self-reported location strings, where available) and their network information, in a hierarchical manner. We further empirically confirm our notion that geo-tagged tweets are not particularly informative when seeking to classify users' home locations, as is done in many previous publications in this field. Unlike previous geo-locating methods for Twitter users, our method only performs costly API calls for pulling network information when absolutely necessary, otherwise relying on previously stored location string data. Furthermore, we are able to estimate certainty for any individual network-based location estimate derived from the distribution of network features for a given user. Finally, we provide full access to our method in the form of an open-source software package, accessible on GitHub[15]. We think that this method has the potential of significantly improving social science research using Twitter data, as location accounts for a considerable degree of variation in many individual and group-level attributes and processes.

However, it is also important to note that we developed this software for scientific purposes, and never with the intention of using individual-level location data for the purpose of identifying, tracking or monitoring individuals. This includes, but is not limited to, micro-targeting, advertising, political messaging, publication of individuals' data, or any activity not compliant with Twitter's terms of service, applicable local laws, or ethical scientific practice. Specifically for the purpose of using geo-located Twitter user data for social science research, we recommend encrypting potentially sensitive data, and never publishing any information that violates the norms of informed consent. In other words, we strongly encourage anyone intending to use our software to *not* publish or share their data at the user-level.

Apart from introducing our hybrid geo-locating method, we also demonstrate its benefits in a typical political science research scenario employing large-n samples of Twitter data, the geo-spatial dynamics of Twitter-based conversation on the Democratic party 2020 presidential primary. Given that our pipeline attached a location classification to 1,766,739 users, we were able to trace regional differences in attention attributed towards different candidates throughout the United States. Furthermore, by combining these user-level locations with simple metrics of relative attention to candidates at the user-level, we were able to produce a granular, temporally and geographically dis-aggregated analysis of Twitter public opinion. Crucially, we illustrate the difference the unit of analysis can have when studying public opinion through the lens of Twitter. While an analysis which operationalises attention toward candidates in a given state as the number of mentions a candidate receives in that state paints a picture of dominance by Bernie Sanders in the first period of data (Jan-May 2019), the same data at the user-level paints a vastly different picture (see Figure 9). Crucially, this picture is considerably closer to measures of public opinion obtained from offline sources (public opinion polls) than that generated by the tweet-level analysis.

While tracing user-level attention to candidates aims to approximate existing types of public opinion measuring, we also demonstrate a further use for geo-located Twitter

[15] Depending on the publication status of this paper, you may need to request access in order to view and clone this repository. It will eventually be publicly accessible under this link.

data in public opinion research. Our data is akin to a giant panel study, where we can trace the temporal evolution of public opinion at the US state level. This is demonstrated in Figure 4.11, where we trace the candidate mentioning behaviour of users who had predominantly mentioned Kamala Harris in the first period of data collection. We observe a shift towards Bernie Sanders, the more liberal candidate, in more liberal states, and toward Joe Biden, the more conservative candidate, in more conservative states. We suggest that this shows a "reversal to the fundamentals" process, given the narrowing number of options for voters. Furthermore, the replication of our data collection in 2020 shows how the users in our sample have shifted their mentioning behaviour in tandem with media coverage, polling, and past election results.

Unfortunately, our data collection period for the 2020 sample stops on February 12th 2020, i.e. one day after the New Hampshire primary - which Bernie Sanders won. At this point in the campaign, Sanders appeared to be the candidate to beat, while all other candidates apart from Biden were widely written off. However, Biden outperformed expectations in the South Carolina primary, leading the remaining field of candidates to rally behind him, and eventually led him to be the nominee and later president. So, while we have shown that we are able to trace public opinion at the state-level, and that users mentioning political candidates appear to adapt their mentioning behaviour over time, there are a number of questions that remain: Do Twitter users change their candidate mentioning behaviour (attention toward candidates) as a response to media reporting, or polling numbers, or are users changing their mind, thus causing this to be reported in the news, or measured in polls? Our geo-locating method could be employed for such a research project, whereby a similar analysis as the one described in this paper is performed, however with a more granular temporal component.

We suggest that further useful research in this area should be devoted to integrating geo-located user-level Twitter data with other measures of user-level attributes, such as age, gender/sex, ethnicity/race or socio-economic status, thus furthering the understanding of how users express themselves politically, or otherwise, on the platform. While some efforts have been devoted to such estimation techniques, the majority of user-level attributes remain unknown for the majority of users. Furthermore, it is important to stress the importance of sampling when it comes to Twitter-based public opinion research. Our sample, while comprehensive, is not perfect. It contains a lot of users tweeting about Democratic presidential candidates, not a perfectly representative subset of the Democratic electorate. Indeed, there is clearly a lot of noise in our sample, beyond the 28 percent of users not located to the US. If the aim of this type of research is to measure public opinion from Twitter while also applying to the offline world, more will have to be done to learn about the distributions of certain variables and quantities *within* the Twitter population to be able to transpose it to offline target populations. However: having the ability to geo-locate users is a significant, necessary step in the right direction, which allows researchers to provide considerably more nuanced analyses of political behaviour as documented on Twitter than without it.

Appendix A

A.1 U.S. State Level Weights

When performing network-based estimations in Sections 4 and 5 we account for the fact that we are more likely to observe users to be connected to people from more populated U.S. states. We use information about the proportion of U.S. Twitter users who live in each state[1] to weight the state locations of reciprocal friend-followers who self-reported to be in the United States. You can see these weights in the *Prop. U.S. Twitter users* column in Table A.1. In most cases the proportion of the population living in each state is similar to the proportion of Twitter users (e.g. .121 *versus* .122 for California). However, in some cases the difference is of a larger and more substantive magnitude (e.g. .002 *versus* .014 for Washington D.C.). For this reason we decided to weight for the proportion of Twitter users instead of the proportion of the U.S. population.

State	Prop. U.S. Population	Prop. U.S. Twitter users	Population - Twitter
California	0.121	0.122	-0.001
Texas	0.088	0.089	-0.001
Florida	0.065	0.065	0.000
New York	0.061	0.075	-0.014
Pennsylvania	0.039	0.033	0.006
Illinois	0.039	0.040	-0.001
Ohio	0.036	0.035	0.001
Georgia	0.032	0.034	-0.002
North Carolina	0.032	0.023	0.009
Michigan	0.030	0.027	0.003
New Jersey	0.028	0.024	0.004
Virginia	0.026	0.017	0.009
Washington	0.023	0.021	0.002
Arizona	0.022	0.019	0.003
Massachusetts	0.021	0.029	-0.008
Tennessee	0.021	0.023	-0.002
Indiana	0.020	0.019	0.001
Missouri	0.019	0.016	0.003
Maryland	0.018	0.016	0.002
Wisconsin	0.018	0.016	0.002
Colorado	0.017	0.019	-0.002
Minnesota	0.017	0.019	-0.002

[1] To calculate these proportions we used the self-reported Twitter location of about 90,000 random U.S. users. These random users were selected by generating random numeric user IDs and then checking whether the users existed.

State	Prop. U.S. Population	Prop. U.S. Twitter users	Population - Twitter
South Carolina	0.016	0.013	0.003
Alabama	0.015	0.015	0.000
Louisiana	0.014	0.012	0.002
Kentucky	0.014	0.014	0.000
Oregon	0.013	0.014	-0.001
Oklahoma	0.012	0.013	-0.001
Connecticut	0.011	0.010	0.001
Iowa	0.010	0.010	0.000
Utah	0.010	0.008	0.002
Nevada	0.009	0.012	-0.003
Arkansas	0.009	0.009	0.000
Mississippi	0.009	0.007	0.002
Kansas	0.009	0.007	0.002
New Mexico	0.006	0.005	0.001
Nebraska	0.006	0.007	-0.001
West Virginia	0.006	0.009	-0.003
Idaho	0.005	0.003	0.002
Hawaii	0.004	0.005	-0.001
New Hampshire	0.004	0.004	0.000
Maine	0.004	0.004	0.000
Montana	0.003	0.004	-0.001
Rhode Island	0.003	0.004	-0.001
Delaware	0.003	0.003	0.000
South Dakota	0.003	0.002	0.001
North Dakota	0.002	0.002	0.000
Alaska	0.002	0.003	-0.001
Washington, D.C.	0.002	0.014	-0.012
Vermont	0.002	0.003	-0.001
Wyoming	0.002	0.001	0.001

Table A.1: Proportion of the U.S. Population and U.S Twitter users who live in each state.

A.2 Confusion Matrix

True (Cencus−Predicted) State Location

Network−Predicted Location

	Wyoming (0.07,0.50)	Wisconsin (0.62,0.65)	West Virginia (0.39,0.35)	Washington, D.C. (0.26,0.47)	Washington (0.61,0.57)	Virginia (0.60,0.56)	Vermont (0.27,0.62)	Utah (0.52,0.62)	Texas (0.66,0.70)	Tennessee (0.68,0.67)	South Dakota (0.25,0.67)	South Carolina (0.67,0.69)	Rhode Island (0.27,0.68)	Pennsylvania (0.79,0.66)	Oregon (0.62,0.60)	Oklahoma (0.76,0.74)	Ohio (0.80,0.66)	North Dakota (0.28,0.68)	North Carolina (0.74,0.66)	New York (0.76,0.52)	New Mexico (0.36,0.48)	New Jersey (0.75,0.59)	New Hampshire (0.27,0.54)	Nevada (0.49,0.58)	Nebraska (0.71,0.76)	Montana (0.29,0.52)	Missouri (0.70,0.56)	Mississippi (0.58,0.75)	Minnesota (0.79,0.69)	Michigan (0.81,0.69)	Massachusetts (0.79,0.66)	Maryland (0.68,0.63)	Maine (0.27,0.41)	Louisiana (0.64,0.74)	Kentucky (0.77,0.74)	Kansas (0.54,0.68)	Iowa (0.72,0.72)	Indiana (0.73,0.66)	Illinois (0.81,0.62)	Idaho (0.25,0.64)	Hawaii (0.36,0.56)	Georgia (0.62,0.61)	Florida (0.62,0.57)	Delaware (0.22,0.50)	Connecticut (0.59,0.54)	Colorado (0.59,0.53)	California (0.54,0.60)	Arkansas (0.60,0.65)	Arizona (0.63,0.57)	Alaska (0.17,0.36)	Alabama (0.72,0.71)
Alabama	4	2	5	4		1		2	16	10		3	1	3			1					1		2		1		2		4	1	1	1	6	2	1		3	2	1	2	6	9	1			22	1	2	2	332
Alaska				1	2	1		1	2	2					1				1	1				1		1	1	2	1	1				1	1		1		2	2		2	3		1		6			23	
Arizona	5	1	3	8	2		4	8	27	7	4		2	1	4	2	5	2	1	7	1	1	5	7	2	3	1	1	1	3			2			2	3	1	3	3	1	6	16	2		8	42		280	1	1
Arkansas	1	1		1	3	2	1	1	20	2	2	3	1	1		1			6				1	2		1	1	2	1					3	1	1					1	4	4			1	12	170	2	2	4
California	31	15	10	80	28	4	12	18	119	22	15	15	21	12	19	11	17	18	17	44	19	12	14	47	8	15	11	12	11	20	26	13	18	24	10	9	15	10	26	22	22	46	71	7	10	23	1745	6	43	16	12
Colorado	6		2	6	3	2	2	4	24	4	2		1	1		2	2	2	2	5	1	1	1	4	1	1	4		5	1	2	1	3	2	2	4	1	3	13	4		5	10		3	214	35	2	1	2	3
Connecticut	1	2		6	1	2			8	5	1		2		2		3			10		2	2			1		1	1	2	3	2	2					1	1	1		5	4		119	1	24	1	1		
Delaware				1	1				3			1		3			3																					1				1	2	17			1				
Florida	14	11	13	35	4	6	2	10	79	23	6	13	11	10	6	2	12	8	18	30	6	11	10	17	5	6	7	14	7	13	7	8	5	11	4	2	2	5	5	6	7	53	981	6	7	7	129	4	15	11	10
Georgia	7	5	3	17	5	9	3	1	29	12	4	17	3	4	3	5	6	2	18	12	2	1	7	8		1	3	7	1	5	2	6	2	15	2	3	1	13	4	11	3	645	31	3	4	5	52	9	1	2	33
Hawaii	1		1	1	1			1	2			2		1			1			1	1			1	1							1	2					1	1	1	52	1	2			1	8		1	3	2
Idaho	1	1		1	2	1		1	4		1											1				1								1						48			3	1	1	1	4				
Illinois	5	17	3	12	3	2	4	1	27	9	1	5	2	5	1	4	6	2	4	13	2	1	4	5	1	4	9	6	4	10	2	5	2	5	3	4	10	11	612	8	3	23	17	3	6	6	65	4	7	3	4
Indiana	3		3	8		1			18	6	1	1	1	2	1	1	4	2	1	3		1	1			1	2	2		4		1	1	2	7		1	285	2	3	1	7	9	2		3	33	4	2	1	
Iowa	1			7	2				13	2	3		1	1		2				1	3	1	1	1	2	1			1	1	1	3	2			1	204		8				7				9	2			
Kansas		1						1	5			1	1	2	1	3	1	1				2	1		1	1	6	1		2		2	1	2		114	1		2	1	1		1		1	2	5	1	1		
Kentucky	1	1	3	1	2		1		12	4		2	2		1	1	10	1		2			1	2		1	2			1	1		1		293	2	1	3		2	2	5	4	4		2	12	1	1		3
Louisiana	1	1	2	2	1			3	21	3	2	3		2	1				2	1	1			1			2	10	1	1		2	3	287				2		1	1	6	4				8	2	3	3	2
Maine	1		2	1	1	1	1	1	6	1	1		2			1			1	2	1	1	2				2			1	1	1	38				1	1	1	2		4	3				10	1	2		1
Maryland	2			25		5	2	2	10	1		4		9		1	1		2	7	4			2	1	2	1				5	254			1			2	2	3	1	10	11	1	1	1	24	2			3
Massachusetts		5		17	3		2	4	16	1	2	2	11	3	2		2		3	12	1	3	8	3	3	1		2	2	4	436	3	3	1	1	2		1	3	3	1	6	16	2	7	4	46	1			1
Michigan	8	3	3	10	1	2	2	4	39	3	8	1	1		5		6		2	6		2	1	3	2	1	2	8		516	2	2	6	5	2	2		3	4	2	2	8	18	2		4	31		4	1	4
Minnesota		13	1	2	2		2	1	9	4	3	1	1	2		1	3	5		5	1	1	1		1		2	2	276	1		3		1	1	2	7	1	3	1		3	3		1	1	25		1	2	2
Mississippi			2	1		1			10	1		2					1	1	1	1				2				182	1	6				2	1	2	2					4	7			2	1	3	1		2
Missouri	3	1		5	1			1	20	2	3	1	1	1	1	2	2	2		4	1	4		3	2	1	215	2		4		1	2	1	1	29	3	2	6	3	1	6	9		3	4	18	5	2	2	
Montana	1	1	1		5	1			5			1	1		2		1			2		1				38			1						2				1	2			1			1	4		1		
Nebraska	2		1	1					4	1	2	1	1	1		1	1				1				124		1	1	2					2	1	1	2			1			4				3				
Nevada	2	1		4	1	1	1	1	12	2	2	3	1	2	4	2	4	3		6	1	1	1	184					1		2	2	1	1		2		2	2	1		6	16				35	3	2	1	2
New Hampshire	2						2	1	2					1									45	1		1					4		1		1						10		2		1	1	7	1			
New Jersey	8	2	4	10	3	1		2	28	5	2	4	4	12	1			1	3	29		354	3	4	1	1	1			3	2	3	3	2				1	1	2	3	15	15	2	1	2	47	3	6	3	
New Mexico								1	4	1	1					2	2				42	1	1	2		2	1			1				1		1				1		3	6	1		1	7	1	1	1	1
New York	11	6	13	49	12	12	10	11	39	11	9	12	15	23	5	3	18	3	11	903	7	28	14	15	4	10	4	1	7	7	17	10	11	10	7	5	4	1	6	7	9	45	55	4	21	17	178	6	13	12	3
North Carolina	2	1	3	6	2	10		1	22	5	1	10		5		1	2	2	390	5		3	5	2			1	3	2	1	5	3		4	3	1		5	2	1		19	18	2	2	3	26	3	1	1	2
North Dakota				2	1				1			1			1			27											1					2												1			1		1
Ohio	6	5	14	6	1		4	1	26	9	2	4	6	5	4		582		3	7	3	1	2	4	1	3	3	4	1	5	1		5	4	10		4	9	7	5	2	15	25	2		2	47	4	2	2	4
Oklahoma	2	1						1	21	3	1		2	3		242	1			4	2		2		1		3	1				1	1	1	1		1	1	1	1			3	1		4	15		1	1	1
Oregon	3			4	5			1	18	1	1	1	2	1	184	1	1		1	2		2	1	10			2	1		1	1		5	1				1	2	4			9	1		2	28	2	5		1
Pennsylvania	7	5	7	10		4	3	3	39	2	4	2	2	565	1	1	2	1	3	10	3	11	1	5		2	1	5		5	5	6	6	1	3	1	1		3	6	1	15	22	3	3	5	53		6	6	2
Rhode Island									3	1			45							1				1		1					1	2	1		1				1				3				2			1	
South Carolina		1	3	7	1	4			20	2		275	2	1	2	1	2	1	4	3			3	3	3	1		1		1	2	3	1	2	1	1		3	2		1	5	7		1	3	15			2	4
South Dakota		1					1				35						1			1						1			2		1				1		2	1		1			1				2	1			
Tennessee	4	1	5	5		4	2	2	29	426	2	5	2	2			5	3	4	1	1	3	3	4	1	2		15	2	3	1	4	2	6	5	2	3	5	4	3	1	5	13		1	1	27	6	2		5
Texas	20	9	13	38	16	7	4	9	1855	20	13	8	3	10	12	20	9	3	16	26	8	7	15	20	5	11	11	19	6	5	9	7	11	26	5	13	6	10	13	13	5	28	45	4	3	14	137	21	18	10	10
Utah	2		4	2				116	10			1			3		1	2		2					1			1	2	1		2			2		1		1		2	2	6	1	1	4	17		1		
Vermont							25			1	1		1							1			2								1													1		1	5				
Virginia		2	11	30	1	219	1		8	5	1	2	5	3		2	6	1	3	7		3	3	2	2		2		1	2	4	5	1	1	5		1	1		2	2	10	7	1	1	2	18		3	1	1
Washington	4	5	4	2	217	1	1	2	13	1	1		2		13	2	4	3	1	4	2	1	1	3		3	3			1	2			4		1	2	1	1	2	2	3	8	1	1	1	47	1	4		
Washington, D.C.	3	1	4	166	19	9		3	15		2	3		5	5	1			2	4	1	4		3		5			3	3	1	11	1	2		1	2	2	1	2	2	6	5		2		40	3		1	
West Virginia	2		93	10		44			14	4	1	1		4	4	1	2		4	2		3	2	1	1	3		3	1		1	3	1		1	1	1	2	1	2	3	6	11	3			22	3	1		2
Wisconsin		199		3	1	4	1	2	16	2	1		3	2	1	1		2		4				2			2	2	5	2		1	1	4		1	2	1	4	1	1	5	5			3	15		1		2
Wyoming	14						1		1					1			1		1																			1					2		1		2	1	1	1	

Figure A.1: Confusion matrix showing the network-predicted *versus* actual (census-predicted) state locations for 24,914 U.S. Twitter users. *Note: In the x-axis, the numbers in parenthesis indicate state-level precision and recall*.

A.3 Measuring prediction uncertainty

In Table 5 we described a set of individual network features that could potentially help us predict whether the network method did a good job at estimating the correct location for a given user: *Network Size*, *Mode Size*, *Second Option Size*, *Number Other Options*, *Number Other People*, and *Non-Modal Dispersion*. Then we assessed the performance of 1,974 statistical models that used different combinations of these individual network features to predict whether our network-based state estimates were correct. We observed some models to do much better than others. However, two main questions remain unanswered: a) which of these individual network features are positively *versus* negatively correlated to accuracy? And b) which of them helps us predict accuracy the most?

In Figure A.2 we answer the first of these two questions. The Figure describes the correlation between the set of network features and the accuracy of our network-based estimates. In particular, it shows the distribution of the t-statistic for each of the times a given feature was included in one of the 1,974 logistic regression predicting the accuracy of the method. At the top we observe for example that we are *more* likely to estimate the correct state location when a larger number of a user's reciprocal friends are located in the modal state (*Mode Size*). On the contrary, at the bottom we observe that we are *less* likely to estimate the correct state location when a user's reciprocal friends are located in a larger number of different states (*Number Other Options*).

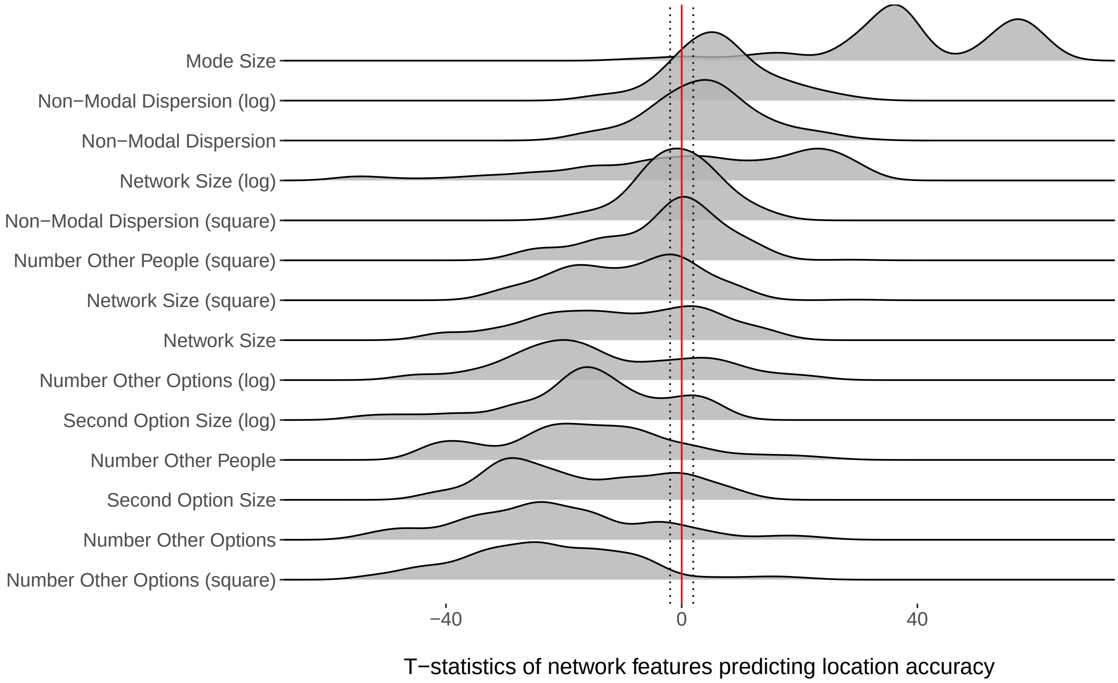


Figure A.2: The relationship between network features and location accuracy.

In Figure A.3 we address the latter question. The Figure describes which of the individual network features turned out to be the most useful at predicting accurate network-based locations: it shows the proportion of the best models that included a particular feature as input. We observe *Mode Size* to be a key feature: 100% of the best 10 and 50 models (based on the highest f-score) used *Mode Size* as input feature. The *Number Other Options* and *Non-Modal*

Dispersion are also useful signals, whereas *Network Size* turned out to be quite uninformative (particularly its raw linear form).

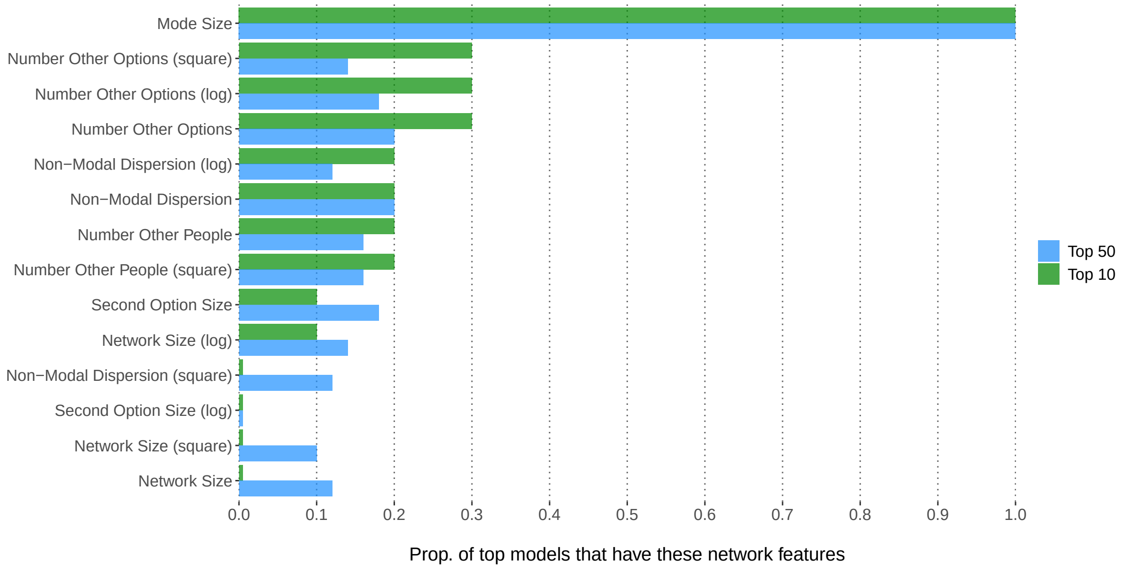


Figure A.3: Network features that contribute to generating accurate state estimates.

Below we see in purple the number of users (out of the 25,778 used for validation) we predicted to be located in each state, whereas in green we observe the ones for which we are at least 60% confident about our predictions.

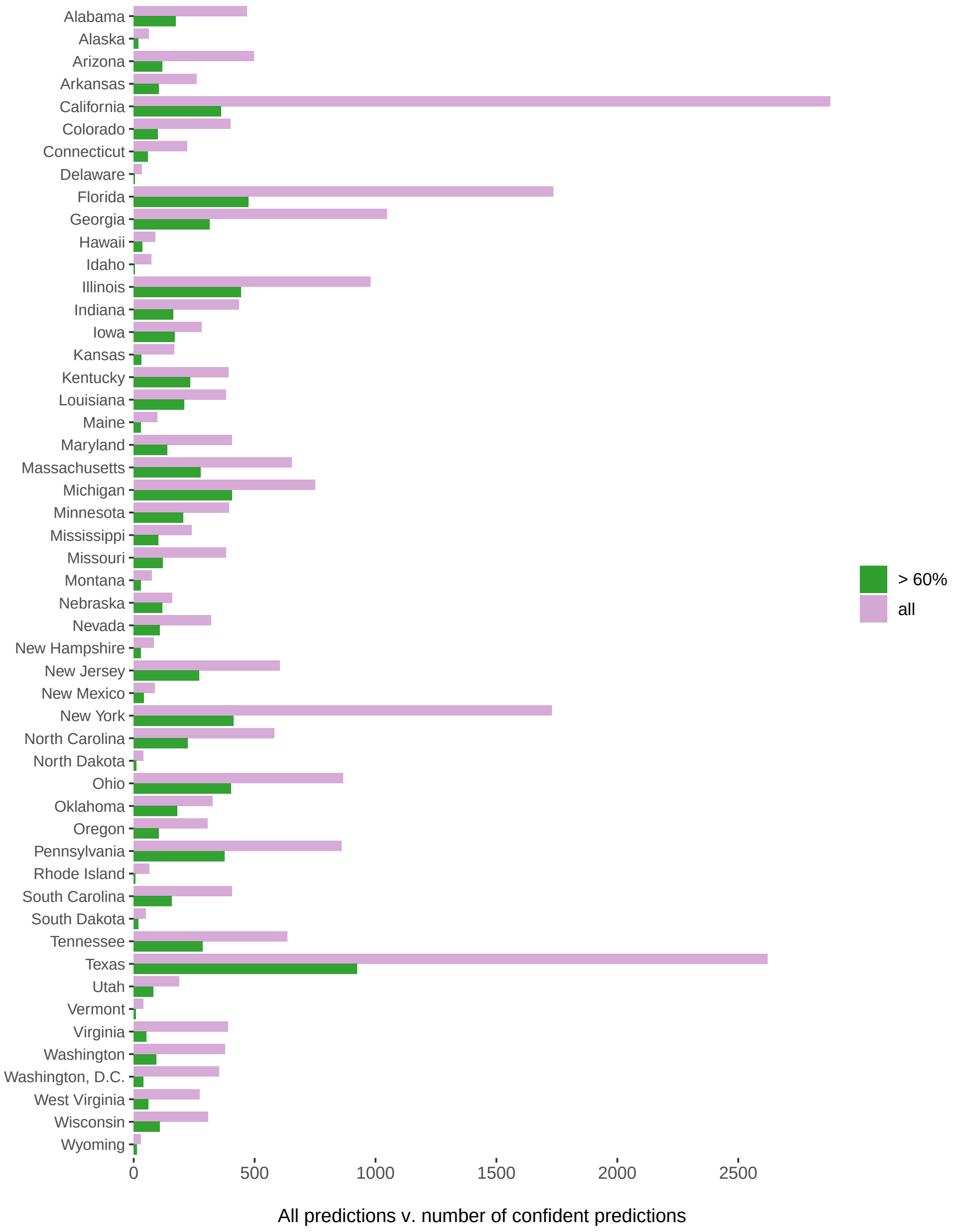


Figure A.4: Comparing the number of users located in each state according to different levels of uncertainty.

A.4 Percentages of users with top candidate

State	Biden	Buttigieg	Harris	Klobuchar	Sanders	Warren	Yang	Steyer
Alabama	28.46	2.88	26.73	0.38	25.00	16.35	0.19	0.00
Alaska	25.98	2.36	18.90	0.00	28.35	24.41	0.00	0.00
Arizona	32.73	6.60	19.37	0.17	23.65	17.05	0.43	0.00
Arkansas	28.29	8.86	20.29	0.00	23.71	18.86	0.00	0.00
California	22.70	4.71	24.91	0.23	27.61	19.22	0.60	0.01
Colorado	23.68	6.41	17.37	0.10	25.75	26.37	0.31	0.00
Connecticut	21.82	7.68	20.81	0.20	25.86	23.43	0.20	0.00
Delaware	35.37	8.54	8.54	0.00	21.95	24.39	1.22	0.00
Florida	36.58	3.38	23.18	0.28	24.64	11.71	0.22	0.00
Georgia	24.82	4.20	31.91	0.17	19.02	19.59	0.28	0.00
Hawaii	22.97	9.46	20.95	0.00	23.65	22.30	0.68	0.00
Idaho	33.33	9.03	15.28	1.39	25.00	15.97	0.00	0.00
Illinois	22.70	7.75	20.15	0.12	22.83	26.16	0.25	0.04
Indiana	18.21	38.53	11.97	0.10	15.69	15.19	0.30	0.00
Iowa	20.86	8.39	12.26	1.29	23.01	33.33	0.65	0.22
Kansas	29.20	8.85	13.72	0.88	27.43	19.91	0.00	0.00
Kentucky	30.91	5.45	21.09	0.18	22.36	19.27	0.73	0.00
Louisiana	24.82	3.31	28.13	0.18	25.74	17.28	0.37	0.18
Maine	26.96	4.41	16.67	0.00	23.04	28.43	0.00	0.49
Maryland	22.30	3.75	30.99	0.09	22.12	20.57	0.09	0.09
Massachusetts	16.86	5.22	10.40	0.24	16.29	50.81	0.19	0.00
Michigan	26.77	6.59	20.72	0.16	24.25	21.19	0.31	0.00
Minnesota	19.31	5.83	14.99	12.73	21.57	25.46	0.11	0.00
Mississippi	27.24	3.11	26.46	0.39	23.35	19.46	0.00	0.00
Missouri	25.13	6.84	19.87	0.26	20.79	26.71	0.39	0.00
Montana	31.34	1.49	17.91	0.00	29.10	20.15	0.00	0.00
Nebraska	21.36	6.36	18.64	0.45	31.36	21.36	0.45	0.00
Nevada	27.32	3.85	23.12	0.00	26.80	18.04	0.88	0.00
New Hampshire	16.80	8.20	19.67	1.64	27.05	26.64	0.00	0.00
New Jersey	26.47	3.53	23.24	0.22	27.50	18.46	0.44	0.15
New Mexico	26.22	5.62	16.48	0.00	28.46	23.22	0.00	0.00
New York	22.18	5.06	19.59	0.22	25.59	27.05	0.28	0.03
North Carolina	22.17	4.81	25.85	0.09	22.45	24.34	0.28	0.00
North Dakota	43.75	3.13	4.69	6.25	17.19	23.44	1.56	0.00
Ohio	29.95	4.44	20.77	0.17	23.49	20.66	0.52	0.00
Oklahoma	25.50	6.49	19.02	0.00	27.74	20.58	0.67	0.00
Oregon	19.98	4.86	16.78	0.11	25.72	32.34	0.22	0.00
Pennsylvania	31.43	3.62	19.57	0.18	22.46	22.46	0.27	0.00
Rhode Island	25.79	4.21	16.32	0.00	28.95	24.74	0.00	0.00
South Carolina	29.86	6.53	25.97	0.47	21.62	15.40	0.00	0.16
South Dakota	26.67	6.67	16.67	3.33	26.67	20.00	0.00	0.00
Tennessee	32.18	3.89	22.35	0.54	21.60	19.01	0.43	0.00
Texas	39.27	2.14	26.44	0.14	20.52	11.29	0.15	0.05
Utah	23.56	5.75	13.79	0.29	22.13	34.20	0.29	0.00
Vermont	14.37	5.75	10.34	0.00	51.72	17.82	0.00	0.00
Virginia	27.19	4.61	24.00	0.24	19.50	23.76	0.71	0.00
Washington	19.97	4.51	17.52	0.46	24.71	32.75	0.08	0.00
West Virginia	28.46	5.85	19.10	0.39	27.29	18.91	0.00	0.00
Wisconsin	24.73	5.60	15.86	0.62	25.19	27.99	0.00	0.00
Wyoming	31.15	6.56	22.95	0.00	24.59	14.75	0.00	0.00
Washington, D.C.	24.31	6.61	18.88	0.81	21.10	28.20	0.05	0.05

Table A.2: State-level percentages of users for whom candidate makes up more than 50% of overall mentions

Chapter 5

Understanding Political Sentiment: Using Twitter to map the 2016 Democratic primaries

Abstract

Sentiment Analysis is a widespread technique for computing the emotional loading of text; to determine if it is "positive", "negative" or "neutral". Sentiment analysis has received broad adaptation in Twitter-based public opinion research, as it provides a framework for approximating support towards a concept of interest, such as a political party or candidate. However, we argue that much of the use of sentiment analysis, specifically in research aimed at forecasting election results using Twitter data, has been prone to a number of questionable design decisions which may contribute to the mixed track record of sentiment-assisted Twitter-based election forecasting. Crucially, previous publications (1) make no distinction between linguistic and political sentiment, thus potentially distorting their measurement validity, (2) typically analyse data at the tweet rather than the user-level, thus biasing results towards prolific tweeters and (3) treat relevant "negative" tweets as uninformative.

In this paper, we provide an overview on different forms of sentiment analysis, their uses and previous applications specifically in Twitter-based election forecasting. Then, we conduct a comprehensive empirical analysis in a novel, three-fold case study approach applying the same methodologies towards predicting three separate elections in the US 2016 Democratic presidential primary in New Hampshire, South Carolina and Massachusetts. Besides replicating previously applied methodologies, we expand the sentiment analysis for Twitter-based public opinion research toolkit with a method for ordinal, intensity-focused political sentiment classification, and further develop a modelling approach which incorporates negative-classified information. We present twelve vote share prediction models for all three primaries.

We find that weighting Twitter data for computationally inferred user-level characteristics, such as home location and political affinity improved sentiment-based vote share prediction accuracy, but find that the inclusion of negative tweets does not consistently improve analyses. Furthermore, we find that shifting analyses from the tweet to the user-level benefits resulting predictions.

5.1 Introduction

The rise of social media services such as Facebook and Twitter has brought with it a new form of *consequential data* (see e.g. Purdam and Elliot, 2015, for a discussion of data classes). These data have some advantages over data conventionally used in the social sciences: costs for data collection are drastically reduced, and potential sample sizes are considerably larger. Furthermore, when regarding such data as containing signals pertaining to public opinion, they offer a potential advantage over *intentional data* such as those derived from surveys: interviewer effects, social desirability bias and other artifacts of the researcher-administered stimuli used to generate these data are absent as the data are created voluntarily.

There is a growing body of research which has used such data to address empirical questions, in fields as varied as Public Health, where tweets were used to predict the spread of swine flu (Ritterman et al., 2009), Criminology, whereby scholars used Twitter data to build models dynamically predicting the occurrence of violent behaviour (Wang et al., 2012), Economics, using tweets in stock market development forecasting models (Si et al., 2013), or Geology, where twitter data were used to develop dynamic earthquake tracking models (Sakaki et al., 2010). In political science, the use of these data has been particularly prevalent; be it investigating twitter users' ideology (Barberá, 2015; Bond and Messing, 2015), the effect that exposure to political tweets has on individuals' political knowledge (Munger et al., 2016) or the phenomenon of group organisation for political protests using social media (Theocharis et al., 2015).

A sub-field of political science using such consequential, or 'digital trace data', uses data obtained from Twitter to forecast election results. Early on and somewhat prematurely Tumasjan et al. (2010) proclaimed to have matched the predictive accuracy of polls by merely counting mentions of parties on Twitter, but Jungherr et al. (2012) replicated this study with divergent findings[1], arguing against the feasibility of forecasting real-world, offline events using mere tweet volume. Other scholars have attempted to forecast elections with Twitter data, but as yet there is no "unifying theory" or "standard approach" in the literature.

For a large amount of Twitter-based public opinion research, specifically that which aims to predict manifest measures of public opinion such as vote shares in democratic elections, *sentiment analysis* has proven a valuable tool for maximising the value of such research. In its simplest form, sentiment-assisted election forecasting assumes that positive sentiment expressed toward a political candidate or party on Twitter (and other social media sites) is a more robust predictor of eventual vote share than mention counts for said candidate or party. While this may be intuitively plausible, computational sentiment analysis, whether it is lexicon-based or employs machine learning, often falls short of adequately classifying sentiment in tweets which relate to a politically salient topic. The key problem is that much of the language prevalent therein is not easily classifiable at scale: for instance, consistently identifying sarcasm and evolving, event-led context-specific language is at best complex and at worst intractable for automated sentiment classification, regardless of the approach used.

In this paper, we investigate areas of interest at the intersection of tweet-based sentiment and Twitter-based public opinion research. Our goal is to understand how insights might be best extracted from user-generated social web text for public opinion research. Our focus

[1] Jungherr et al. (2012) replicated Tumasjan et al.'s study with a completely different sample of tweets, with neither paper providing any information on the process employed to collect this data, illustrating another issue with research in this area.

is primarily methodological. In essence, this means that we apply established steps in the sentiment analysis toolkit, as well as novel alternatives and outline the workflow for moving from data to insights. Crucially, we then collate these insights into a tangible empirical application - the election vote share estimation paradigm - in order to evaluate the efficacy of research design decisions in this sub-field, with a particular focus on weighing up the costs and benefits of design choices in such applications.

This line of enquiry involves four core areas of investigation. First, we outline existing approaches to measuring sentiment in political tweets, and emphasise the distinction between *political* and *linguistic* sentiment. We also introduce our approach to measuring political sentiment in tweets, which involves hand-annotation of samples of pertinent tweets in a multi-step process: a categorical addressee annotation, followed by an ordinal sentiment annotation. Second, and relatedly, we discuss the up and downsides of measuring sentiment at the categorical versus ordinal level, and empirically investigate how an ordinal measure can enhance the resulting predictions/aggregations. Third, we trial different approaches of aggregating sentiment scores in order to model vote share percentages in an election. We compare a novel methodology - simulating votes by randomly drawing sentiment scores from samples of sentiment-labelled tweets - with established methodologies. Fourth, we investigate how sampling decisions - from the definition of a geographical sampling frame, to the selection of labelling sets, to the weighting and de-biasing of samples for analysis impacts the model results.

In order to apply these aspects of our investigation to an empirical context, we selected the 2016 US Democratic Party Presidential primaries as our case study. Specifically, we chose three primary elections; New Hampshire, South Carolina and Massachusetts. We opted for this case selection, as the US primary system allows us to study cases with different underlying socio-demographic "fundamentals", while still having the same electoral rules and participating candidates. This allows for both high-precision evaluations of a given predictive model's configuration, but also a robust way of getting towards the core aim of this project - learning more about which conditions - be they arising from contextual factors of a test case or the configuration of a model - impact the efficacy of using Twitter data for public opinion research.

In the remaining parts of this paper, we begin by reviewing the existing methods for systematically extracting subjective opinions - *sentiment* - from text taken from the social web[2]. We describe the various approaches; their respective advantages, shortcomings and limits. We then discuss the literature on Twitter-based election forecasts and public opinion research which is supported to some degree by sentiment analysis. We assess the methodologies which were employed to produce a given forecast as well as their respective performance records. Then, we describe research design and the data used in this project, outlining how samples were generated, and how the samples for the three different states differ in terms of geographical and ideological makeup. We then describe our hand-labelled *political* sentiment annotation procedure. We then outline how we propagated ordinal sentiment scores from labelling sets to larger samples of pertinent tweets. Finally, we describe the performance of twelve predictive models for estimating vote share percentages in the 2016 Democratic presidential primaries in New Hampshire, South Carolina and Massachusetts, and draw

[2] By using the term 'social web' rather than 'social media posts' or even 'tweets', we include all user-generated text on the web, rather than just that posted on popular platforms such as Twitter or Facebook.

conclusions both regarding best practice of sentiment-assisted Twitter-based public opinion research, as well as the efficacy of the endeavour as a whole.

5.2 Sentiment Analysis of social web texts

Social web text is as an abundant resource of user-generated data containing opinions on every conceivable topic. Using computational methods in order to analyse and provide quantitative estimates of the emotional content of those opinions ("X is good, Y is annoying") within text units is typically referred to as *sentiment analysis*.

Following Liu (2012), we concur that textual sentiment can be represented using two measures, *polarity* (is a given text unit expressing positive, negative or neutral sentiment?) and *intensity*[3] (*how* positive or negative is the sentiment). Understanding and categorising sentiment - at the very least in terms of polarity - is a relatively simple undertaking for the average human. However, human judgement is not sufficiently scaleable for large samples of web text. Hence, researchers typically use computational tools to generate estimates of the sentiment polarity and intensity.

Initial attempts focused on matching words to semantic lexica, such as specific emotions (Huettner and Subasic, 2000). Pang et al. (2002) provided an early example, succeeding in estimating sentiment polarity with an accuracy exceeding that of "human-produced baselines" (p. 79), assigning polarity estimates for a large sample of online, user-generated movie reviews, reaching accuracy levels of up to 82.9 % (p. 83) using several different machine learning (ML) approaches. Such approaches use labelled training data[4] by which the algorithm "learns" both sides of the estimation equation. This "knowledge" is then applied to unlabelled data. In 2003, Nasukawa and Yi (2003) introduced a lexicon-based approach to estimating sentiment polarity for "web pages and news articles" (p. 70), achieving similarly high levels of estimation accuracy to Pang et al. (2002). The lexicon-based approach differs significantly from the ML approach, in that a pre-defined reference lexicon containing a large number of individual words is labelled with either a polarity indicator or an ordinal intensity score. A basic model of lexicon-based sentiment would simply look for the occurrence of any such word in a given text unit, and combine polarity/intensity ratings of all occurring words to produce a sentiment estimate. More complex models take into account a word's position within a sentence, and so on.

Several researchers produced papers with a similar goal of estimating sentiment of online texts, typically reviews of products or movies (Turney, 2002; Dave et al., 2003; Hu and Liu, 2004; Pang and Lee, 2004; Liu et al., 2005; Popescu and Etzioni, 2007, see e.g.) . Gamon et al. (2005) introduced an ML approach using supervised support vector machines capable of analysing the sentiment of single sentences with "precision scores ranging from 0.95 to 0.97" (p. 10). Despite high levels of precision achieved in the publications listed above, none of these papers estimate sentiment *intensity*. This endeavour, was first tackled in the literature by Pang and Lee (2005). They introduced a method capable of precisely estimating review ratings using Support Vector Machine regression. Taboada et al. (2011) provide a

[3] This is referred to as *sentiment rating* by (Liu, 2012, p. 30). We choose the term "intensity" instead of Liu's "rating" as the latter is specifically relevant to user-produced ratings of product or movie reviews but less so to political sentiment.

[4] in this case, the unit of analysis was individual movie reviews, with user-generated ordinal ratings attached to them, thus eliminating the need for time-consuming labelling by the researchers.

different solution to this problem, building a sentiment analysis tool that estimates polarity *and* intensity on an ordinal scale relying on a reference lexicon rather than a ML approach. The authors assert that this approach outperforms ML algorithms for domain-independent data. The accuracy of their method ranges from 65 % to 81.5 %, depending on the nature of the dictionary and the data. Finally, Thelwall et al. (2012) developed an application - "SentiStrength" - which measures sentiment intensity and polarity by matching text units to a lexicon. However, intensity is expressed on two ordinal scales, one positive and one negative. The authors argue that this approach allows for the possibility "that both positive and negative sentiment can coexist within texts" (p. 164).

Both the ML and lexicon-based approaches have advantages and shortcomings. While ML sentiment analysis is likely only applicable to the topic domain of the text units for which labelled data, and thus training data exists, it has the potential for outperforming lexicon-based methods (see e.g. Thelwall et al., 2012, p. 172) by capturing indirect sentiment indicators, such as terms which convey sentiment in the context of the topic domain of the training set, without necessarily having the same connotations in general language. Lexicon-based methods will likely not capture these terms, and they will only work for languages in which lexica are available. However, lexica do not require newly labelled data for each new project, and tools can be re-used for applications with text data covering widely disparate topics. Hence, the evidence suggests that for general applications, e.g. with text not clearly situated within one distinctive topic domain, lexicon-based methods are likely more appropriate, whereas ML approaches are best suited to narrow, domain-specific tasks. This assessment is supported by Ribeiro et al. (2016), whose "SentiBench" project analyses different "state-of-the-practice" sentiment analysis tools with the aim of comparing their respective output for types of text typically mined for sentiment (social media posts, reviews and website comments). They find that "there is no single method that always achieves the best prediction performance" (p. 3), and "existing methods vary widely regarding their agreement" (p. 3).

Specific application of the sentiment analysis paradigm to tweets has been widespread. This is hardly surprising, given the simplicity of collecting user-generated text data from the platform, their wide use across academic and commercial settings and the structural simplicity of the tweet as a text unit mean that method and data are well mapped.

Pak and Paroubek (2010) attempted to replicate the state of the art of ML approaches for classifying sentiment on a corpora of tweets. They find that a configuration using text tokenised to bi-grams using a multinomial Naive Bayes classification algorithm produces best results when producing categorical (i.e. polarity) ratings for tweet-level sentiment. Agarwal et al. (2011) showcase a similar experiment, however with the added output of two lexica, denoting key tweet features mapping emoticons to sentiment polarity, as well as "an acronym dictionary collected from the web with English translations of over 5000 frequently used acronyms" (p. 31). In addition to using hand-annotated data to train their sentiment detection algorithm, they incorporate their emoticon-dictionary and a reference sentiment lexicon into their analysis, resulting in a hybrid model. In essence, the authors map each token in a tweet to one of fifty pre-defined features, allowing them to then estimate a sentiment polarity for a given tweet. The authors use data collected straight from Twitter's public sample stream - implying that their method is aimed at classifying "generic" tweets of no particular topic domain. They achieve maximum accuracy levels of around 0.75 (p. 36).

A similar approach was taken by Zhang et al. (2011), whereby a lexicon-based approach is utilised in the first step of the analysis, and is effectively used to annotate tweets for sentiment polarity. These annotations are then used to train a Support Vector Machine classification algorithm providing this classification pipeline with accuracy levels exceeding the existing state of the art. Crucially, their hybrid approach outperforms all established methods against which they benchmark it, yielding maximum accuracy levels of 0.85 (p.7).

Abbasi et al. (2014) perform a similar sentiment benchmarking analysis to Ribeiro et al. (2016), however exclusively using twitter data with content relating to specific topic domains. They find that different tools are more accurate in different topic domains, and the best overall performance was achieved by "SentiStrength", with an average accuracy level of 0.67 (p.5). Severyn and Moschitti (2015) take the Twitter-based sentiment classification paradigm into the state-of-the art of Machine Learning, by classifying tweets in a three-step approach using convolutional neural networks and distant supervision. The authors state that their model is among the best in the Semeval-2015 challenge for advancing the field of Sentiment Analysis. This model achieves a maximum accuracy of 0.85 (p. 962) on a dataset of tweets, classifying sentiment polarity.

5.3 Sentiment Analysis and Twitter-based election forecasts

Studying, understanding and predicting public opinion using online text data - especially in the context of voting behaviour and its outcomes - has received considerable scholarly interest in recent years. Using user-generated online text as data offers several advantages over intentional survey and polling data: costs for data access and collection (at least in the case of Twitter, easily the most popular data source in this field) are drastically lower than with survey-based approaches; data are available in large quantities and sentiment conveyed within tweet text is a self-generated expression of opinion rather than a narrow response structured by the parameters of a survey instrument.

The sub-field of election forecasting/vote share prediction by means of data collected from social media sites began with Tumasjan et al. (2010), who predicted the outcome of the 2009 German Bundestag elections with a level of accuracy comparable to that of pre-election polls by counting mentions of relevant parties and candidates on Twitter. While some scholars (Gayo-Avello, 2013, e.g.) argue that this was merely down to luck, and a more clearly specified approach would have resulted in different - inaccurate - results (Jungherr et al., 2012), several publications have subsequently attempted to forecast election outcomes using Twitter data, with varying degrees of methodological rigour, theoretical foundation and model performance.

While some researchers employ an approach similar to that of Tumasjan et al. (2010) (Gayo-Avello, 2011; Metaxas et al., 2011; Jungherr et al., 2012; DiGrazia et al., 2013; Caldarelli et al., 2014, see e.g.), with a mixed record of accuracy, a large amount of the Twitter-based election forecasting literature uses sentiment analysis. O'Connor et al. (2010) found that lexicon-derived tweet sentiment correlates highly, ranging from r=.77 to r=.81 (p. 128), to opinion poll time series on US presidential job approval[5] when aggregated over long time periods. While the authors note that this task is inherently easier than that of forecasting the actual *outcome* of an election, their findings form a crucial building block and justification for further research into the area, as it highlights that Twitter sentiment on a political topic

[5] for US President Barack Obama in 2009

closely mirrors sentiment elicited from individuals through the traditional survey paradigm.

Building on this foundational work by O'Connor et al, several scholars conducted published research attempting to take these findings a step further, and actually use insights derived from tweets through sentiment analysis in order to forecast the outcome of democratic elections. Sang and Bos (2012) used machine learning to classify tweet sentiment, coupled with mention counts to forecast the Dutch Senate elections of 2011. The authors achieved an adequate level of predictive accuracy - the country's leading polls were more accurate by four seats. However, the authors used polling weights after the fact to stratify their model. Bermingham and Smeaton (2011) used a similar approach, but with a significantly worse performance: a mention count model outperforms their positive sentiment volume model. However, this is arguably due to specific research design weaknesses: a small sample of labelled tweets, low levels of inter-annotator agreement and a significantly lower number of tweets overall compared to similar publications. Ceron et al. (2014) used perhaps the most sophisticated sentiment-assisted method for their forecast of the French Presidential and legislative elections of 2012, using a supervised ML approach adopted from Hopkins and King (2010) to adequately forecast the winner of the Presidential race within the error range of traditional polls as well as the vote share of most parties in the National Assembly[6]. Ceron et al. (2015) replicated this approach using different case studies (The 2012 U.S. presidential election and the Italian Democratic Party's 2012 primary), achieving results that outperformed polls in the majority of cases[7]. Finally, Burnap et al. (2016) used Thelwall et al. (2012) "SentiStrength" lexicon-based sentiment tool to forecast the 2015 UK general election, which broadly matched pre-election polling averages, but failed to forecast the Conservative party's eventual victory.

Overall, the evidence to date allows for four tentative conclusions:

1. Sentiment analysis-based approaches to predicting/forecasting metrics related to electoral outcomes outperform models relying solely on candidate/party mention-volume (however, rigorously applied sentiment analysis and a large enough sample are necessary conditions for this to be the case).

2. Machine learning-based methods have a higher predictive accuracy than lexicon-based methods (especially using advanced algorithms, such as Hopkins' and King's (2010), rather than simple approaches such as Naive Bayes).

3. While basic research such as Severyn and Moschitti (2015) suggests that advanced machine learning approaches (such as convolutional neural networks) can improve tweet sentiment classification accuracy and potentially reduce the need for costly human annotation of domain-specific tweet data for model training, these findings are based on tweets from the general domain, not topic-specific tweets.

4. There appears to be no unifying theory for *why* Twitter-derived, sentiment-based predictive public opinion models perform well or fail. The existing literature takes little

[6] The authors' algorithm performed well for the large parties, PS and UMP (Ceron et al., 2014, p. 350), but over-estimated support for left-leaning parties while under-estimating support for right-wing parties. This finding ties in to evidence from studies on Twitter's demographic composition (Rainie, 2012; Smith and Anderson, 2018), suggesting that left-leaning and young individuals are over-represented on the platform

[7] In addition to forecasting the popular vote, the authors forecast swing state vote shares, whereby the poll of polls outperformed the Hopkins and King (2010) method in only two out of 11 cases (Ceron et al., 2014, p. 10)

> note of either conceptual or methodological factors that may contribute to the outcome of predictions. Conceptually, there is little discussion of why precisely measured sentiment in tweets mentioning politicians should tell us something about that politician's electoral chances. Methodologically, existing research is agnostic to interventions such as weighting, de-biasing; the whole question of sampling; or the dearth of user-level demographics.

In the following section, we delineate the key issues that persist within the sentiment-based public opinion modelling literature as a whole. Following this, we outline our approach to measuring the *political* sentiment within tweets containing politically salient text, and our methodology for predicting electoral vote share proportions. We use sentiment-annotated tweets in the three empirical case studies of the 2016 Democratic presidential primaries in New Hampshire, South Carolina and Massachusetts.

5.4 On the relationship between political tweeting and voting

Overall, the idea of modelling election results using data gathered from Twitter or other social media sites has been widely critiqued, see e.g. Lui et al. (2011); Jungherr et al. (2012); Jungherr (2015); Gayo-Avello (2011, 2013). Gayo-Avello (2013) in particular points out the structural problems inherent with Twitter data as predictors of political behaviour: the demographics of the "Twittersphere" do not match those of target offline populations, and the dearth of available information on users' socio-demographic attributes (e.g. age, gender, race/ethnicity, socio-economic status, etc.) makes drawing externally valid inferences complex at best and impossible at worst. Furthermore, Twitter data are not robust to self-selection bias (Gayo-Avello, 2013, p. 670) - tweeting is a voluntary act, meaning that tweet collections do not contain a sample of political opinions of Twitter users, but rather a sample of political opinions of those users choosing to share them. Furthermore, Jungherr (2017) points out an often overlooked factor that shapes most any collection of Twitter data used for academic research. Rather than selecting a sample of tweets pertinent to a given topic of interest by sampling from the population of *all* existing tweets or users, tweet collections are typically compiled by filtering Twitter's streaming API based on *a priori* defined keywords. This method is subject to two key shortcomings: First, there is no guarantee that researchers correctly specified relevant keywords before the fact, meaning that the collection may be over- or under-sampling from the population of all political salient tweets. Secondly, the fact that Twitter limits the proportion of tweets that are accessible through their streaming API, and further, that there is no guarantee that these tweets are representative of the population (of interest), means that any sample of tweets obtained from the streaming API may be biased in unknowable ways. Jungherr refers to this phenomenon as the "n=all fallacy". However, short of a costly subscription to Twitter's full stream, there is little researchers can do to overcome these shortcomings.

These are top-level/structural issues associated with social science research using social media data. However, some of these issues can be mitigated. Most promisingly, this can be done in estimating user-level socio-demographic attributes, where several researchers in the field have developed computational tools that estimate such attributes, for instance users' age (Nguyen et al., 2013), their political affinity (Barberá, 2015) or their home location

(Loynes et al., *forthcoming*. We argue that incorporating such measures of individual-level user attributes into any analysis of political tweets can be useful, and thus do so in this paper.

Notwithstanding the complex systemic issues associated with using tweets for public opinion research, applying sentiment analysis in such a setting is further prone to bias originating from imprecise operationalisation of the theorised causal behavioral relationship between posting social web texts and engaging in (offline) political behaviour, such as voting. Researchers typically equate the proportion of positive tweets mentioning a party/candidate in a corpus of salient tweets with eventual vote share percentages in elections. This assumes an underlying model of proportionality between positive sentiment expressed on Twitter and individual-level vote choice. However, researchers do not typically explain their assumptions regarding a) the appropriateness of sentiment analysis in twitter-based election forecasts, and b) what positive sentiment in tweets signifies; in other words, what they are actually *measuring*. Furthermore, tweet-level analyses overweight active posters, essentially giving them multiple "votes" in such models. Hence, we argue that it is of importance to downsample to the user-level, thus providing a degree of proportionality which mirrors the "one person, one vote" maxim in democratic elections.

Crucially, there are three core questions that need to be asked concerning the relationship between political tweets and (offline) political behaviour. First, there is the fundamental question of how political tweeting and political behaviour are linked; what common origins they share and if and how we can transfer meaning from one domain to another. Second, there is the more specific question of what it means when a given Twitter user tweets something with positive political sentiment about a politician or party. Third, there is also the question of what it signifies when a user tweets something negative about a politician.

5.4.1 The linkage between political tweets and political behaviour

There is an implicit assumption prevalent in the literature on sentiment-based public opinion research using Twitter data: positive sentiment in a tweet directed toward a given party or candidate is an adequate proxy of voting intention for said party/candidate. It is an intuitive hypothesis that highly popular candidates typically poll better than unpopular candidates, and eventually receive more votes (Lewis-Beck and Rice, 1982; Kenney and Rice, 1983). O'Connor et al. (2010) found a high correlation between longitudinal positive sentiment in tweets mentioning political figures and their survey-derived approval ratings. The authors' findings are plausible - tweeting one's opinions on a politician is similar to answering a survey question on one's views towards them. A similar hypothesis seems appropriate for the sentiment volume / vote share relationship: candidates for whom we measure low amounts of positive sentiment on Twitter should be expected to get fewer votes than those with a larger volume of positive sentiment.

It is clear that there is *no* direct causal relationship between political tweeting and political behaviour, such as voting. However, as displayed in Figure 5.1, some theoretically sound assumptions can be made regarding their relationship. Firstly, certain latent (and potentially observable) individual-level factors and attributes shape any individual's behaviour - including their propensity to vote, or to tweet. When it comes to political tweets, it is further apparent that many such factors likely exert an equal, shared or subsequent influence on both types of observable behaviour. While one may speculate as to whether these factors might be stable, like partisanship or ideology, or more acute and unstable, such as an evaluation

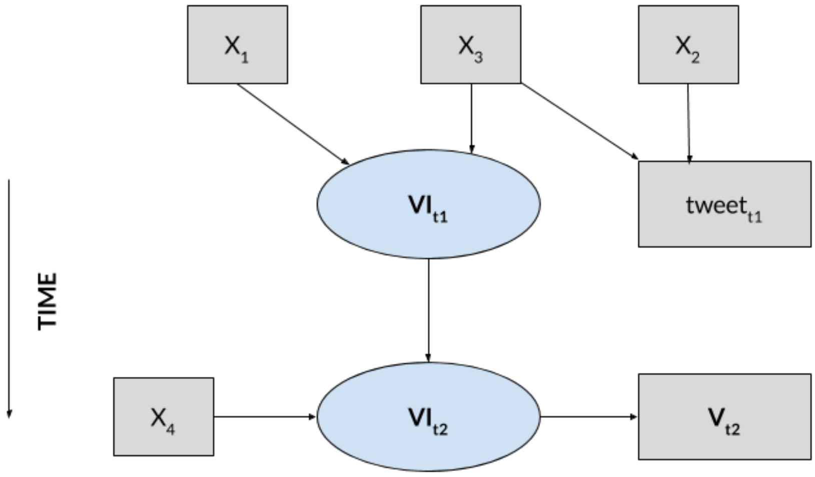


Figure 5.1: A general model of the connection between vote choice and political tweets. **VI**: Voting Intention. **X**: Latent factors. **V**: Vote.

of a politician's response to exogenous events, e.g. a war or a pandemic (Mueller, 1970; Newman and Forcehimes, 2010, see e.g.), or something subject to medium-term evolution, such as an individual's evaluation of the state of the economy (Lewis-Beck and Stegmaier, 2007, see e.g.), it is highly plausible that such factors contribute to both behaviours of interest. However, it is conversely also clear that there are certain factors which may affect these two behaviours in opposing ways (such as the age or geographical location of an individual, which may make them likelier to vote, but significantly less likely to be present or active on Twitter (Smith and Anderson, 2018), while there are other factors which likely share no common influence over the two.

Hence, we conclude that further research is necessary in order to establish the core factors determining online and offline political behaviour, in order to better model the latter using the former. Our core argument here however is to a) establish that there is clearly a strong correlation with offline and online political behaviour at the individual level, and b) that our assessment of the relationship between the two suggests that a conceptualisation of unweighted positive tweet volume as a proxy for vote share percentages is likely a faulty approximation, especially when off-the-shelf tools which measure linguistic rather than political sentiment are employed. Hence we argue for Twitter-based public opinion research using sentiment analysis to a) focus on the user-level rather than the tweet-level and b) measure political rather than linguistic sentiment. We believe that these simple changes have the potential of improving the efficacy of resulting predictive models.

The *meaning* of positive and negative political tweets

Given the assumption that political tweets and offline political behaviour are highly correlated and share causal factors, it is useful to further consider how political sentiment measured therein, be it represented categorically (positive, negative, or neutral) or ordinally (on a scale from X to Y) can be used to model offline political behaviour. In order to address this, it is useful to first assess the way in which existing research has operationalised this, and then investigate what a given sentiment rating for a given tweet will typically signify when transposed to the dimension of the author's relationship to the pertinent issue. In essence, previous forecasts simplify this relationship by assuming that candidate C_i's share of positive tweets will be directly related to their share of the vote share in election E. When expressed in a formal equation, the function of tweeting and voting is as follows:

$$\hat{P_C} = \frac{\sum_{C,t} I(S_{Ct} > 0)}{\sum_t I(S_t > 0)} \tag{5.1}$$

where $\hat{P_C}$ is an estimator for the proportion of votes for candidate C and S_t is the level of sentiment of tweet t.

We suggest that this conceptualisation is somewhat short-sighted in that it does not address the assumed, implied or otherwise, meaning of positive tweets, and further does not address negative or neutral tweets at all. Hence, let us discuss our understanding of what these types of tweets signify when transposed to the offline world:

A positive tweet about a politician likely indicates a positive disposition by the author toward said politician. As we know from existing research, citizens tend to vote for candidates whom they like, whereas candidates who are not broadly liked face an uphill electoral struggle. So, the author has expressed that they feel positively towards the politician. However, this does not mean that they do not like another politician who is competing in the same election. However, oftentimes we can verify if this is the case by also assessing the author's other political tweets, and aggregating the sentiment expressed within them to all pertinent politicians in order to confidently estimate which one they like best. ***A negative tweet about a politician*** conversely most likely indicates a negative disposition by the author toward said politician. Furthermore, this suggests that the tweet's author is less likely to vote for this politician. Again, we can oftentimes assess whether this a firmly held belief or just an *ad-hoc* reaction that may deviate from the author's overall opinion by looking at all the author's relevant tweets.

Now, when it comes to using this information in modelling a given candidate's vote share, it intuitively makes sense to tally tweets they have received which are positive, and compare them versus all opponents. This is how previous sentiment-assisted Twitter-based election forecasts operated. While there are core issues with this methodology which we have discussed above, such as analysing the data at the tweet rather than the user-level and thus likely amplifying the most frequent tweeters, who are also likely to be prolific partisans (Conover et al., 2012; Barberá et al., 2015), or neglecting the inherently skewed nature of a given sample of tweets, there is a further implicit assumption in these models, namely that tweets with negative sentiment toward a politician or party are uninformative for such a predictive model.

5.4.2 Tentative conclusions and research agenda

We have introduced the core concepts of the literature of sentiment-assisted Twitter-based election forecasts and have provided an in-depth interpretation of the underlying implicit models of tweets and votes employed in previous publications. Overall, this allows us to draw the following tentative conclusions, leading us to a research agenda for our implementation of sentiment-assisted Twitter-based public opinion prediction:

1. It is unclear what precisely tweet-level positive sentiment volume of politically salient tweets measures. Evidence suggests that it is highly correlated with approval ratings derived from opinion polling, but this correlation is likely contingent on large samples of data in order to alleviate the issues arising from a tweet-level analysis, as well as the noise inherent in such data.

2. This understanding has led researchers to use positive sentiment volume as a proxy for vote choice rather than for approval. We argue that this design choice may account for some of the unreliability in previous forecasts, and it may be useful to incorporate a transposition from candidate (dis)approval to voting intention in models.

3. Previous research has widely ignored tweets with negative sentiment in their analyses. We suggest that negative tweets may be just as informative as positive tweets, and their incorporation into predictive models should hence be explored.

4. Tweet-level analyses are prone to bias due to oversampling prolific partisans and should be replaced with user-level analyses.

5. Off-the-shelf tools for sentiment analysis (especially lexicon-based ones) are likely not suitable for correctly estimating *political* sentiment in political tweets, but rather focus on linguistic sentiment, which in turn is likely not capturing certain innate elements of a given tweet's political sentiment. Twitter-based public opinion research relying on sentiment-derived measures should seek to measure political sentiment if possible.

6. While certain variables are hard to adjust for when working with samples of Twitter data, such as a propensity to turn out to vote, others are estimateable with state-of-the-art tools. Incorporating such user-level estimates (e.g. location and political affinity) as a way of weighting tweet samples may be a fruitful avenue for improving predictive accuracy.

5.5 Research Design

We now introduce our approach to extracting sentiment analysis from samples of politically salient tweets from three distinct geographical locations, the US states of New Hampshire, South Carolina and Massachusetts, from three weeks before their respective Democratic party presidential primaries in 2016. Crucially, we generate political sentiment scores at the tweet-level by hand-annotating sub-samples of pertinent tweets. We outline coding rules resulting in tweet-level ordinal scores expressing what we refer to as "political sentiment". This score incorporates both sentiment *polarity* and *intensity*, and transposes the underlying political meaning of a tweeter's expressed political sentiment toward an addressee (one of the two candidates in all of the primaries, Hillary Clinton or Bernie Sanders), into an ordinal score of political sentiment. We further propagate these scores to larger samples of pertinent tweets using machine learning.

We use these political sentiment scores in twelve models to predict the candidates' vote shares. In these models, we implement the research agenda outlined in the previous section, which allows us to draw conclusions as to the appropriateness of the undertaking as a whole, as well as the usefulness of individual models. Besides evaluating the efficacy of models' aggregation and weighting approaches, we also contrast the usefulness of using smaller, exclusively hand-labelled samples versus larger samples of machine-annotated data.

Besides the established approach of comparing positive sentiment volume for the different candidates competing in an election, we also showcase a novel methodology to twitter-based election forecasting which aggregates hand-annotated political sentiment scores to vote share percentage estimates by simulating new data from the probability density function of labelled sentiment scores, thus incorporating *all* labelled tweets into the analysis, rather than just those understood to be positive.

5.5.1 Case selection and Data

We focus on three elections within the US presidential primary, the Democratic party 2016 primaries in New Hampshire, South Carolina and Massachusetts. We believe trialling this method on several, smaller elections is a more useful pursuit than testing it on one national-level election, for numerous reasons. First, all three elections share the same candidates, providing comparability and consistency across all data to be sampled and labelled, while all samples can be created from the same source collection. Second, all three elections operate under very similar electoral rules, with minor differences, which are a) known and b) can be accounted for. Third, while candidates remain the same across elections, socio-demographic distributions in these states are different, but known - at least at the aggregate level. The same applies to electoral rules in the different states; New Hampshire has a semi-open system (allowing voters to re-register their party allegiance on the day), Massachusetts has a closed system (allowing only registered Democrats to participate, and registration has to occur over a month ahead of election date) and South Carolina has an open system (allowing anyone to participate). Having these similar test cases with differences along known parameters provides for a useful scenario in seeking to understand model performance in estimating vote share from Twitter data.

	NH	**SC**	**MA**
Primary Date	09 Feb 2016	20 Feb 2016	01 March 2016
n tweets	18,082	33,027	92,104
n unique users	2,061	5,338	13,498
Mean n(tweets) / user	8.9	6.9	9.5
Median n(tweets) / user	1	1	1
Max n(tweets) / user	1024	780	1970

Table 5.1: Summary statistics for geo-located samples of primary-relevant tweets for New Hampshire, South Carolina and Massachusetts.

In order to generate samples of tweets relevant to these elections, we filter our source collection[8] for tweets mentioning the candidates running in these elections, Hillary Clinton and Bernie Sanders[9]. These tweets were collected over the entire course of the 2016 electoral season using Twitter's freely accessible streaming API. Then, we geo-located the unique users in this sub-collection using the method outlined in Loynes et al. (forthcoming), with the goal of providing a municipality-level estimate for each user. We discarded tweets by users who were not unambiguously placed in one of the three states, or whose tweets did not occur within a three-week window leading up to the respective state's primary. This leaves us with three sub-samples of tweets, from which we drew a sample of n=1000 each, whereby we sampled unique users (thus allowing only one tweet by each unique user to be contained within the sample), giving priority to tweets closer to the election date[10]. These sub-samples of n=1000 were then hand-annotated for political sentiment using the guidelines described in the following section by the lead author of this article.

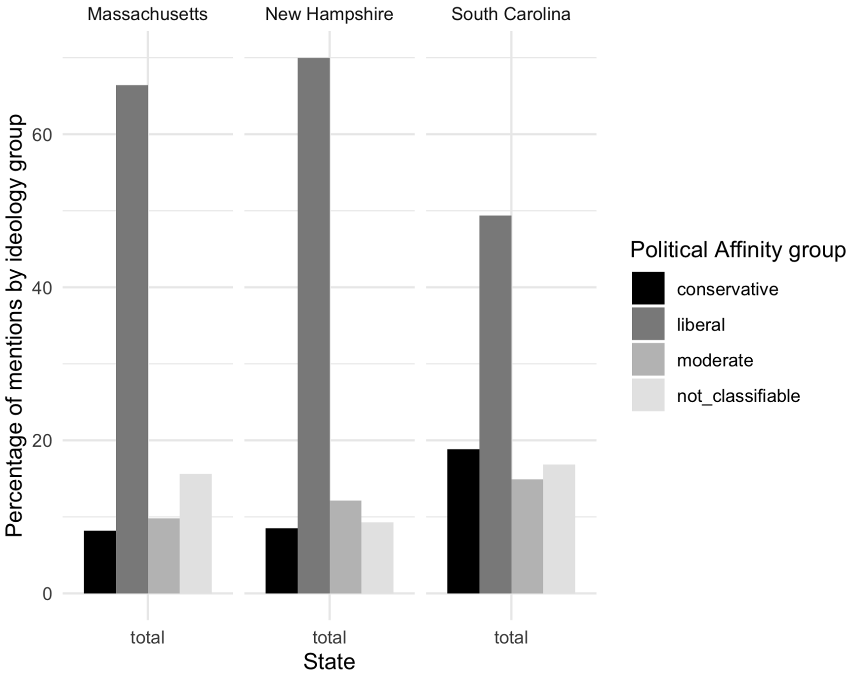


Figure 5.2: Number of tweets in sample, by state and political affinity group

We computed users' political affinity on a two-dimensional left-right scale using the method outlined in Barbera, 2015. We did this in order to be able to perform exploratory weighting on the state-level samples. Users' political affinity is modelled as a function of the news-media and elite Twitter accounts they follow. If, for instance, user u follows three right-wing politicians and two right-wing media outlets, this user will be classified as right-wing, and so on. Figure 5.2 shows the distribution of these affinity classifications, at the tweet-level.

[8] The source collection was a Twitter Streaming API collection beginning at the end of 2015, with all names of candidates running in the election, their official Twitter handles, related hashtags and common (mis)spellings of their names

[9] By "mentioning", we mean not only @-mentions, where a tweet directly notifies the mentioned account holder; but rather all tweets containing a range of words that match the candidates' names.

[10] In other words, if user u tweeted five times in our sample, their most *recent* tweet will be the likeliest to be contained in the n=1000 sub-sample

For all three states, tweets by users classified as liberal dominate our samples. We determine affinity group by matching user-level estimates to estimates for reference accounts. We classify users as *liberal* if their score is equal to or below that of the Washington Post, *moderate* if their score is between that of the Post and the Wall Street Journal, and *conservative* if it is equal to or greater than that of the Journal. It is also apparent that proportionally, the South Carolina sample contains more conservatives than New Hampshire and Massachusetts, where the political affinity distributions (in our sample) are broadly similar. We suggest that this may be explainable in part by the open nature of the South Carolina primary leading to higher participation of opposite partisans in Twitter-based conversation, but is also likely a reflection of the demographics of South Carolina, which is widely understood to be more conservative than the two New England states. However, Figure 4 confirms our intuition regarding the ideological split of people who tweet about the Democratic primary - they are overwhelmingly liberal.

As we computed location classifications for every user in our sample, it is useful to compare how geographically dispersed our set of located users is versus the state's actual population distribution. This is useful as we know that the type of conurbation an individual lives in - a big city or a small village - correlates highly with many factors known to influence voting behaviour (see e.g. Scala and Johnson, 2017). For this purpose, we calculated quartiles of population numbers by municipalities as a proxy for the urban-rural distribution, both for the census-derived real-world distribution and for the distribution of user-level location classifications in our Twitter user samples. These are depicted in Table 2.

State	Type	Q1	Q2	Q3	Q4
New Hampshire	real	3.06	8.21	18.04	70.69
New Hampshire	sample	0.77	1.69	3.32	94.23
South Carolina	real	0.82	3.09	10.38	85.34
South Carolina	sample	0.20	0.56	2.09	97.14
Massachusetts	real	2.21	8.97	19.73	68.93
Massachusetts	sample	0.30	2.04	4.72	92.95

Table 5.2: Percent of inhabitants/users in population-size quartiles, real/sample

Across the states, it is apparent that quartile 4 - the one containing the most populous areas in the state[11] is significantly over-represented in all the samples. This is in line with previously reported, survey-derived findings stating that the Twitter population skews heavily toward users who live in urban centres (Rainie, 2012; Smith and Anderson, 2018). Further, it is important to note that the low proportion of users living in less populated municipalities, especially in more rural states such as New Hampshire and South Carolina, is likely to affect any kind of estimates derived from these data.

5.5.2 Labelling tweets for political sentiment

We now outline the guidelines for generating the most (internally and externally) consistent, and thus accurate hand-annotated political sentiment scores for our three sub-samples of n=1000 tweets. For the purpose of understanding the choices we made, it is useful to first

[11] As the quartiles are calculated using intra-state data, Q4 does not provide inter-state comparability in terms of population size.

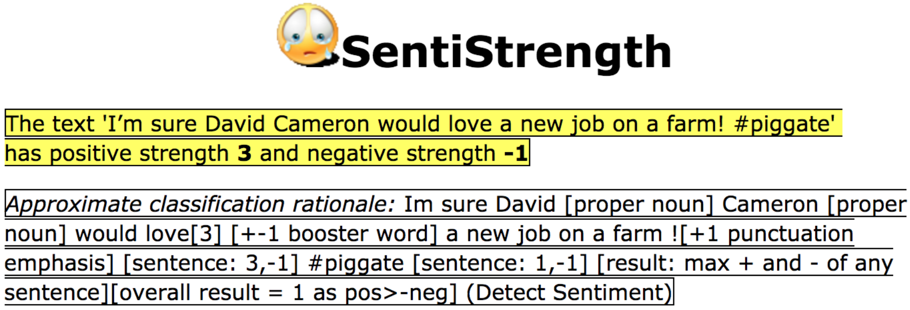
discuss the potential issues arising when using established methods to extract sentiment from tweets.

The Problem: Extracting *political* sentiment from tweets

Consider the following tweet:

"I'm sure David Cameron would love a new job on a farm! #piggate"

While this was not taken from the Twitter stream, it is representative of several of the language attributes prominent in politically salient tweets. First, it refers to a high-profile politician[12]. Second, it refers to an event - #piggate[13] - with specific language applicable only to the context of this event. Third, the author employs sarcasm/satire as a stylistic device - people who are aware of '#piggate' will know that David Cameron is not likely to have wanted to work on a farm. In other words - a reader aware of the political context surrounding Cameron would intuitively understand that this tweet does *not* convey positive *political* sentiment toward him, even though its superficial *linguistic* sentiment may suggest so. However, a lexicon-based sentiment analysis algorithm would grade this tweet positively - as is the case with "SentiStrength" (see Figure **??**), the most accurate sentiment analysis tool (on average) for tweet-specific sentiment (Abbasi et al., 2014, p. 5). Further, an ML classifier trained on generic tweets to detect linguistic sentiment would likely do the same. We argue that this is the core problem of using off-the-shelf sentiment analysis for politically salient tweets.


SentiStrength

The text 'I'm sure David Cameron would love a new job on a farm! #piggate' has positive strength **3** and negative strength **-1**

Approximate classification rationale: Im sure David [proper noun] Cameron [proper noun] would love[3] [+-1 booster word] a new job on a farm ![+1 punctuation emphasis] [sentence: 3,-1] #piggate [sentence: 1,-1] [result: max + and - of any sentence][overall result = 1 as pos>-neg] (Detect Sentiment)

Figure 5.3: Output from the SentiStrength interface for example tweet

In order to alleviate the potential problems associated with extracting politically salient meaning from user-generated social web text, we propose a granular, domain-specific human-coding framework. We present samples of tweets to coders which:

1. were authored by users who live in a target geographic entity.

2. are highly likely to deal with the specific election or other politically salient topic we are interested in learning more about. We achieve this by filtering the Twitter stream for relevant keywords.

[12] the former UK Prime Minister David Cameron

[13] See https://en.wikipedia.org/wiki/Piggate for more information on this particular event

The coders tasked with labelling our tweet samples for political sentiment receive a specific set of instructions on how to label each tweet. In essence, this concise guidance on how to identify five elements in a given unit of text to label is employed:

1. Who is the **addressee** of the tweet. Is it candidate/party A or B, both, or neither?

2. What is the the **polarity** (positive, negative, or neutral) of the political sentiment conveyed in the tweet?

3. What is the **intensity** of the political sentiment conveyed in the tweet? Strong emotive language connotes high intensity, in both polarities. Nonetheless, a soberly worded appraisal of a given addressee which delineates the author's strong (dis)agreement with the tweet's subject can also convey strong intensity. It is important for coders to re-visit previously labelled tweets in order to establish an internally coherent scale.

4. Which **scale** to use? We opt for an 11-point ordinal scale ranging from -5 to +5, whereby 0 indicates neutral. We see this as a fruitful trade-off between labelling complexity and label granularity.

5. What about the **tweet's author**? Is there any indication in the user's profile that may aid our interpretation of their tweet, such as their profile image, profile bio, or other tweets on their timeline?

We see numerous benefits in this approach: the granularity of scores tells us something about how strongly people feel, rather than simply indicating positive versus negative. We can still however use these labelled data as training sets for categorical classification tasks.

It is important to note that this highly involved method of human-coding entails significantly more effort than polarity-only human-labelling tasks for typical ML sentiment analysis applications. Besides merely reading a unit of text and deciding between three response categories, our approach requires a (sometimes complex) multi-step process with a more involved reading not only of the text unit at hand, but also of associated information where available and necessary. Overall, our method of extracting political sentiment scores from social web text aims to avoid the biases inherent in both of the established methods of sentiment analysis frequently used in computational social science research. We see our coding approach as addressing many of these biases.

5.5.3 Machine-propagation of political sentiment scores to larger samples of tweets

We are keen to assess the efficacy and efficiency of using larger, machine-propagated samples of electorally relevant tweets for the goal of mapping offline political opinion, as we suspect that smaller, entirely hand-annotated samples may produce equally useful results. Hence, for the purpose of empirically testing this intuition, as well as to replicate previously employed methods, we use our tweets (n=1000 per state), hand-annotated with political sentiment scores as training data in order to machine-label equivalent scores on the larger samples of pertinent, geo-located tweets (see Table 1). Given our operationalisation of political sentiment, this is a two-step classification/regression problem. First, we need to determine the addressee of a given tweet, after which we can regress an ordinal political sentiment score onto it.

Classifying tweets' addressees

There are two potential avenues for classifying a given tweet's addressee in the context of this research design. First, what we call the "naive addressee classification" approach, which simply filters the text of a given tweet for the names of the relevant candidates, in this case 'Hillary Clinton' and 'Bernie Sanders' (as well as alternative spellings, typing errors, nicknames and official Twitter handles). The possible outcome labels of this classification mechanism are 'Hillary Clinton', 'Bernie Sanders', 'both', and 'neither'.

However, given our previous discussion of how lexicon-based methods (which this one is, albeit with a much simpler remit) may be insufficient for tasks aimed at extracting political sentiment, we also consider a machine learning method. In this case, we take the n=1000 hand-annotated tweets as training data, as addressee labels have been human-annotated at the tweet-level. Then, we train multiple classifiers (Random Forest, Linear Support Vector Classifier, Multinomial Naive Bayes and Logistic Regression) using the scikit-learn library for Python 3.7 (Pedregosa et al., 2011), and evaluate their performance. Across all three states, the Linear Support Vector Classifier performed best, reaching accuracy scores of 0.98 (New Hampshire), 0.98 (South Carolina) and 0.98 (Massachusetts) in 10-fold cross validation using re-balanced training data (i.e. the training data was re-sampled in order to equally contain each possible addressee class).

In the pursuit of the most accurate model, we inspected the overlap between the naive and the machine learning addressee classification methods, by first ascertaining the proportion of cases in which both methods produce the same addressee label. For New Hampshire, **77.4%** of cases shared the same label, while the proportion for South Carolina was **87.6%** and **85.3%** for Massachusetts. Then, we subset each training set to only contain those tweets where both approaches diverged, and inspected each resulting sample by hand in order to ascertain which of the approaches yielded the most valid and reliable results. In all three cases, we found that the 'naive', keyword-filtering approach vastly outperformed the machine learning approach when blindly re-annotating previously mismatched cases and comparing them to the two classification outputs. Hence, we chose to employ the naive method on the larger tweet samples and forgo machine classification for the addressee variable for all three states. This shows that, while lexicon-based methods may not be suitable for classifying political sentiment in political tweets, they are certainly useful for classifying the less ambiguous attributes of tweets.

Regressing political sentiment scores to unlabelled tweets

We further outline our approach to predicting political sentiment scores for all sampled geo-located tweets for the three states (New Hampshire, n=18,082; South Carolina, n=33,027; Massachusetts, n=92,104). This process is more complex than that of classifying addressee-labels to tweets, as models now need to learn numerical text features associated with hand-labelled sentiment scores and use them to predict the same in previously unseen data. As our goal is to have rich, granular score data to use in modelling public opinion (i.e. an ordinal intensity/polarity score versus just a categorical polarity score), we choose to conceptualise this as a regression problem rather than a classification problem. While our ordinal scores are not technically scaled as interval or ratio variables (typically seen as the scale requirements for performing regression tasks in machine learning), we argue that the quasi-metric nature of our scores makes this the preferable option, over e.g. a multi-class or multi-iteration classification approach.

Given our discussion of the intricate, evolving, context-specific language associated with politically salient tweets, we argue that it is prudent to build regressors at the per-state, per-addressee level. While this may be withholding potentially relevant text features from a given model, it means our models are well calibrated to the intricacies of each unique context we are studying. Hence, we build 6 individual regression models - one for each candidate, for each state.

Before we begin training models, we pre-process all text for a given scenario. This involves the removal of stopwords, as well as non-informative Twitter-specific text and symbols (such as '@', 'RT', '#'). Further, we lemmatise all words in a given corpus, meaning that words are reduced to their shortest form which still reflects its core meaning (e.g. 'campaigning' is reduced to 'campaign', and 'well' is reduced to 'good'). These pre-processing steps were undertaken using the 'spaCy' library for Python 3.7 (spaCy, 2020).

In order to build the best possible model for each scenario, we train five different models for each: Linear Regression, Binary Logistic Regression (one versus rest), Logistic Regression (multinomial), Ordered Logistic Regression and Random Forest using the scikit-learn library for Python 3.7 (Pedregosa et al., 2011). We evaluate model performance using the Mean Absolute Error metric (how far is a predicted score from a hand-labelled score, on average). Table 3 shows the mean absolute error of all models, obtained from running 10-fold cross-validation.

State	**NH**		**SC**		**MA**	
Candidate	Clinton	Sanders	Clinton	Sanders	Clinton	Sanders
Model						
Linear Regression	1.95	1.6	1.47	1.73	1.79	1.57
Logistic Regression (one v rest)	2.15	1.89	1.83	1.99	2.19	1.67
Logistic Regression (multinomial)	2.03	1.86	1.79	2	2.13	1.65
Ordered Logistic Regression	1.85	1.56	1.66	1.71	1.85	1.58
Random Forest	1.9	1.69	1.47	1.79	1.85	1.59

Table 5.3: Performance of 5 different regressor models (Mean absolute error), computed through 10-fold cross-validation

It is immediately apparent that the accuracy range of all models is fairly narrow, ranging from an MAE of 1.47 to 2.19. Given a scale ranging from -5 to +5, this means that any model is prone to falsely predicting a given tweet's polarity if the intensity is comparably low. Conversely however, the best models show encouraging performance, with an acceptable error range. While certain models achieve marginally higher performance, we chose to centre our analysis on two models, the Ordered Logistic Regression and Random Forest models, as their performance is most consistent. Hence, we discard the remaining models, and propagate sentiment scores to the three larger samples of tweets for both models.

However, we are further interested to see which of the two models is more useful for our application and hence to employ the best possible predicted values in our analysis. For this purpose, we generated a mean score out of both the predicted Ordered Logistic Regression and Random Forest scores for each tweet. Then, we randomly sampled a further fifty tweets per sample (total n=300), and, keeping the predicted scores blind to the human annotator (this paper's first author), annotated them for political sentiment, using the same rules and scale as in the original training data generation. This exercise found that, across all samples, the Ordered Logistic Regression method provided the most accurate scores, in fact producing exact matches with hand-annotated scores in 38% of cases, while providing an MAE of 1 across samples. Hence, we employ the scores produced by Ordered Logistic Regression throughout our empirical analysis. However, it is important to note that this second stage of model validation and calibration showed that our model typically over-estimated the positivity and under-estimated the negativity of tweets addressed to Bernie Sanders, while showing the inverse for tweets addressed to Hillary Clinton - over-emphasising negativity and under-estimating positivity. We found this to be the case across all 6 Ordered Logistic Regression models.

5.5.4 Modelling vote share percentages

Having described our approach to sampling, political sentiment annotation and machine propagation of both addressee and sentiment labels across larger samples, we now outline our proposed methods to modelling vote share percentages from measured and predicted political sentiment in election-relevant tweets for the 2016 Democratic party presidential primaries in New Hampshire, South Carolina and Massachusetts. Our vote share percentage estimation models can be divided into two broad categories: replication of established strategies and a novel approach. Furthermore, we introduce novel ways of weighting and post-stratifying samples using computationally estimated, individual-level user attributes. Overall, this adds up to 12 models applied to each of the three states.

Replication of established models

As we discussed in "Sentiment Analysis and Twitter-based election forecasts", there exists a considerable amount of literature seeking to estimate vote share percentages in elections by aggregating sentiment scores from relevant tweets, all of which apply a similar methodology, whereby positively labelled tweets are aggregated for each candidate. We replicate this method for all three states. The predicted vote share for a given candidate is modelled as the percentage of positive tweets for them out of all positive tweets toward candidates in a given sample.

Simulating votes from the candidate-level distribution of sentiment scores

In addition to replicating the established method, we introduce a novel method of modelling candidate-level vote share percentages in elections. In this method, we seek to address a core omission in previous related research, namely the assumption that tweets with negative sentiment are uninformative to modelling vote share.

We achieve this by using the probability density function (distribution) of a given candidate's political sentiment scores, and randomly sampling (simulating) new scores from this distribution. We repeat this for each candidate, until we reach an n equivalent to that of the number of expected or actual participants in the target election for all candidates. Hence, we have a dataset with n(rows)=n(voters), and 2 columns, each representing a "simulated approval score" for a given simulated voter. By doing this, we aim to approximate n randomly selected voters, who each have a measure of (dis)approval for either candidate. Then, we model voting intention as a choice between several candidates, with the main contributing factor for vote choice V in election E being an individual voter's (dis)approval of each candidate (R) relative to the others.

Following this, our basic model of vote choice is as follows:

$$V = C_1 \text{ if } R(C_1) > max(R(C_2, ..., C_i)) \tag{5.2}$$

For a two-candidate election, this can also be expressed in terms of the marginal rating R_M, whereby p(V) is constituted as follows:

$$R_M = R(C_1) - R(C_2) \tag{5.3}$$

$$if R_M > 0, V = C_1 \tag{5.4}$$

$$if R_M < 0, V = C_2 \tag{5.5}$$

Using this model, we generate a simulated vote choice variable for each simulated voter in our simulated data. We predict vote share percentages for each candidate by aggregating vote choice variables and calculating the percentage received by either candidate.

Weighting and post-stratification

Given the individual-level user attributes we computed for the data used in this project, we run all our models using different underlying data configurations.

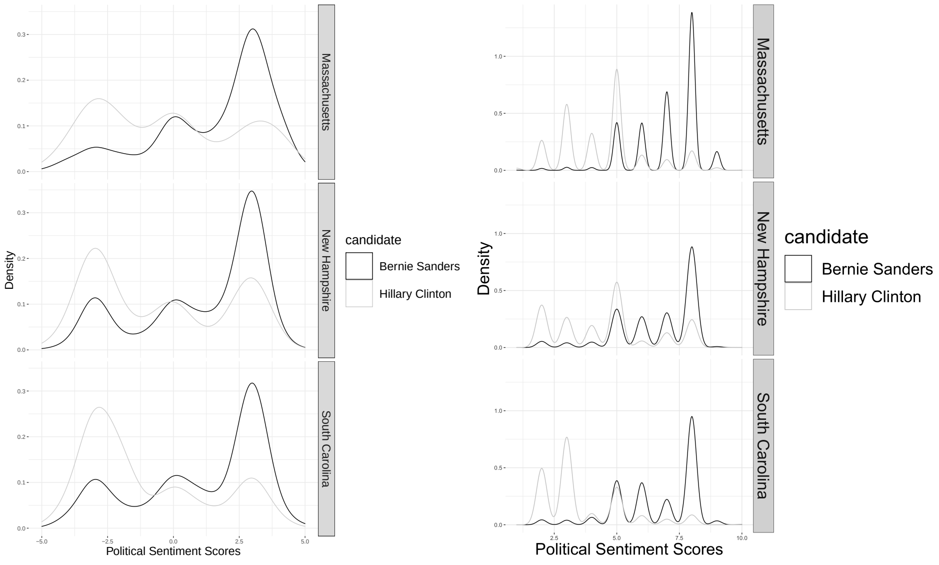
First, we run models at both the raw **tweet-level** and at the **user-level**. In order to achieve this, we subset samples for unique Twitter user-ids and average the sentiment they have expressed toward candidates in all of their tweets which appear in our samples. This post-stratification step has the core aim of reducing the potential distortion caused by prolific partisans who repeatedly tweet similar messages in support (or opposition) of a particular candidate. As voters in democratic elections usually only have the use of one vote, we argue that this should make vote share estimates more predictive of real-world outcomes.

Second, we run our models at the scope of both the **large, machine-labelled samples** and the **smaller, human-annotated samples**. Given our skepticism of the benefits of employing large scales of machine-annotated data when aiming to predict vote share percentages from sentiment-labelled tweets, and further, given our considered approach to sampling training datasets, we believe it prudent to examine the effect of using larger samples albeit with less reliable sentiment scores.

Third, we weight samples by **location distribution**. As a given user's machine-classified home location is one of two inclusion criteria for the data underlying this research, we choose to further employ the granular information contained within this user-level attribute. Specifically, we choose to expand our analysis by re-running models with samples weighted for population distribution. We undertake this weighting step as it is clear that our samples significantly over-represent urban areas, while under-representing rural ones - both of which are individual-level attributes known to be correlated with political behaviour. By over-sampling rural users and under-sampling urban users, we seek to correct for this bias.

Fourth, we weight samples by user-level **political affinity**. As the elections we are studying are all under the Democratic party umbrella, and this party is broadly defined as a liberal, social-democratic party rather than a conservative one - which, in some cases actually prevents non-registered Democrats from participating in primaries - we argue for the usefulness of adjusting target samples so as to more accurately reflect the real-world electorate in terms of its political affinity. Hence, we run a range of models with only data from users classified as liberal or moderate by Barbera's (2014) algorithm.

5.6 Findings



(a) Hand-labelled

(b) Machine-labelled

Figure 5.4: Distribution of sentiment scores for candidates across states

Before analysing the results from our 12 models, we first discuss the distributions of political sentiment scores from which these results are derived. We show these distributions in Figure 5.4. It is immediately apparent that for the hand-labelled sets (a), the distributions for either candidate look very similar: Bernie Sanders has a clear peak on the positive side of the distribution, with a smaller bump on the negative side, whereas this is reversed for Hillary Clinton. We see small differences in the height and width of the peaks between states, for both candidates, but overall it is a very similar probability density function for all three states. Given the fact that the tweets underlying these distributions were all labelled by this paper's lead author, we can confirm that this visual representation of the political sentiment of the training data sets was also felt when labelling: Hillary Clinton, regardless of the state, got a lot of abuse and dislike, while Bernie Sanders has a large number of fans saying only positive things about him - again, independent of which of the three states a tweet may have come from.

Sub-figure (b) in Figure 5.4 shows the distribution of Political Sentiment scores for the machine-labelled samples. Given the fact that these labels originated from the sentiment distributions in the training sets, it is no surprise that these distributions look similar to those in (a), albeit with seemingly more granularity.

From the distributions pictured in Figure 5.4, it is immediately apparent that tweets addressed to Bernie Sanders were more likely to be of a positive nature, while tweets addressed to Hillary Clinton were more likely to be negative toward her. This is not necessarily in line with our pre-labelling expectations, namely that Clinton would receive higher levels of positive sentiment, at least in South Carolina. However, these expectations were guided, at least to some degree, by the assumption that tweets localisable at the state-level would typically concern themselves with state-level, electorally salient issues. However, given the experience of hand-labelling the tweets, we suggest that the majority of tweets followed national stories and events more often than local ones. We re-visit this in the discussion section.

	Date	Hillary Clinton	Bernie Sanders
New Hampshire	09 Feb 2016	37.68	**60.14**
South Carolina	27 Feb 2016	**73.44**	26.02
Massachusetts	01 March 2016	**49.73**	48.33

Table 5.4: Election results in the three primaries of interest

Table 4 shows the actual election outcomes in the three states of interest. While Sanders managed to achieve a pronounced victory in New Hampshire, Clinton dominated in South Carolina. Massachusetts showed a neck-and-neck result. Interestingly, we find ourselves with three states with significantly different demographics, and three considerably different results. New Hampshire is mostly rural, white, and not necessarily too liberal, but tends to receive a huge amount of media attention due to its status as the first primary in the calendar. South Carolina is somewhat less rural, but still not particularly urban, with a plurality of black voters, and a large amount of pension-age voters in the Democratic party electorate. Massachusetts is the most urbanised and liberal of the states, featuring many high-income white collar workers. We argue that these demographic fundamentals contribute considerably to the outcomes of these elections, which is in part why we chose to study these particular elections rather than other ones.

	State	NH				SC				MA			
	Method	Sanders	Clinton	MAE		Sanders	Clinton	MAE		Sanders	Clinton	MAE	
1	+ Sent, full set	57.54	42.46	3.69	✓	58.99	41.01	32.70	✗	59.05	40.95	9.75	✗
2	+ Sent, full set, user-level	61.29	38.71	1.09	✓	64.70	35.30	38.41	✗	61.24	38.76	11.94	✗
3	+ Sent, full set, user-level, location-weighted	61.88	38.12	1.09	✓	65.50	34.50	39.21	✗	61.17	38.83	11.87	✗
4	+ Sent, full set, location & affinity-weighted	56.35	43.65	4.88	✓	57.46	42.54	31.17	✗	58.34	41.66	9.04	✗
5	+ Sent, training set	50.93	49.07	10.30	✓	54.79	45.21	28.50	✗	50.99	49.01	1.69	✗
6	+ Sent, training set, location-weighted	51.77	48.23	9.46	✓	56.30	43.70	30.01	✗	50.19	49.81	0.97	✗
7	+ Sent, training set, location & affinity-weighted	50.23	49.77	11.00	✓	52.40	47.60	26.11	✗	50.06	49.94	0.97	✗
8	Simulation, training set	66.97	33.03	5.74	✓	71.39	28.61	45.10	✗	68.58	31.42	19.28	✗
9	Simulation, training set, location & affinity-weighted	60.18	39.82	1.09	✓	64.27	35.73	37.98	✗	61.04	38.96	11.74	✗
10	Simulation, full set	79.07	20.93	17.84	✓	89.36	10.64	63.07	✗	87.85	12.15	38.55	✗
11	Simulation, full set, location-weighted	83.31	16.69	22.08	✓	89.80	10.20	63.51	✗	88.44	11.56	39.14	✗
12	Simulation, full set, location & affinity-weighted	75.22	24.78	13.99	✓	87.40	12.60	61.11	✗	86.41	13.59	37.11	✗
13	Mean	62.89	37.10	1.66	✓	67.70	32.30	41.41	✗	65.28	34.72	15.98	✗

Table 5.5: 12 vote share estimation models for three states, estimated vote share percentages by candidate, mean absolute error for each model by state, was winner predicted correctly?

Table 5 shows the predicted vote shares of our 12 models for each of the three states. Overall, the models are split into those aggregating positive sentiment in the sample for both the machine-labelled and hand-labelled sets, in non-weighted and weighted configurations, and those where vote share predictions are drawn from newly simulated data contingent on the probability density functions of the state-level samples, again in weighted and unweighted configurations. For each model, we report the estimated vote share percentage for a given candidate, as well as the state-level model's mean absolute error (MAE) and a binary indicating whether the model correctly predicted the winner of the election (✓vs ✗). In addition to the 12 models, we also present a mean model, which contains the average of all models, again at the per-state level.

First off, it is apparent that across all models, each one predicted the correct winner - Bernie Sanders - for New Hampshire, while failing to correctly predict the winner for both of the other states. When looking at the predicted vote share percentages and mean absolute error for the different models for New Hampshire and the other states, most notably Massachusetts, it becomes clear that there is a large divergence in model accuracy (measured in error distance of predicted vote share versus actual vote share). For New Hampshire, three models achieve a very low mean absolute error of 1.09 - the two weighted (location, location and affinity) applications of the positive sentiment method on the machine labelled sample, as well as the fully weighted simulation on just the training set. Otherwise, it is clear that the simulation based on the machine-labelled set widely misses the target, with MAE figures ranging from 13.99 to 22.08. Furthermore, the models applying positive sentiment volume on solely the training set also widely miss the mark, with MAEs around 10. Interestingly, however, while both model types are not fit for purpose, it is striking that the positive sentiment volume (training set) models under-estimate Sanders' vote share significantly, while the simulation (full set) considerably over-emphasises it. Even when comparing the different model types on the same source data set, the minimum distance between vote share predictions is at least 8 percentage points. This suggests that neither the established model type - which treats negatively classified tweets as uninformative and discards them - or our novel method, which ascribes equal importance to negative and positive tweets is entirely adequate in being useable in a vote percentage aggregation worklfow. The three New Hampshire models which perform best are the positive sentiment volume model on the machine-labelled set, at the user-level, weighted by user-level location, or unweighted, as well as the simulation model, performed only on the training set, weighted for all available variables. We are pleased with these results and suggest that this indicates that two of our intuitive cases for adjusting samples - namely the user-level and user-level location - are likely useful in improving vote share prediction accuracy. It is also important to note that the other models, especially positive sentiment volume performed only on the training set, or simulation performed on the machine-labelled set fall widely short of a useful prediction for New Hampshire.

The case of South Carolina is considerably less accurate than New Hampshire, as is shown by an average MAE of 41.41 across all twelve models, and the fact that none of the models correctly predicts the eventual winner. Interestingly, we find that here the least inaccurate model prediction comes from the fully weighted positive sentiment volume model using only training data, a stark contrast with the New Hampshire model, where this is one of the less accurate models. While we will further explore reasons for model performance in the next section, we suggest that this clearly shows two things: if the 'right' (i.e. close to representative) users do not exist in a given geographic entity on Twitter, then no amount post-stratification will make their opinions appear in a Twitter-based study of offline public opinion, and further, the underlying data, its quality and scope may be more important than a given model in order to extract public-opinion relevant insights from them.

The case of Massachusetts provides both cause for encouragement and disappointment, when assessing the various models' performance. First off, the state features the two models with the lowest MAE (both 0.97, positive sentiment volume on training sets, weighted) of any of the models in this study, which however fail to correctly predict the winner of the election. While this is partly due to the fact that the election's actual outcome was extremely narrow, it is nonetheless concerning that even a highly accurate model does not push Hillary Clinton over the 50%-threshold. However, when consulting the distribution of hand and machine-labelled scores in Figure 4, this is not particularly surprising, as the source data for any of these models simply paint a very clear picture, which is not in favour of Hillary Clinton. The other models show a mixed bag of predictions, whereby the worst models are hugely wide off the mark (namely the simulation on machine-labelled samples), indeed for any of the three states. But, overall it is important to note that the results for Massachusetts are not as inaccurate as South Carolina but also not as accurate as New Hampshire. This is potentially a reflection of the states' differing demographics, and can be seen as encouraging going forward, in that this can be understood as a finding in its own right.

Finally, it is apparent that models produce highly similar predictions across states when consulting Table 5 at the row-level rather than the column-level. This is undoubtedly a reflection of how similar the source data for each three states were in regard to their distribution of sentiment scores, and thus again highlights the primacy of data over model in regard to producing the most accurate predictions.

5.7 Discussion

We now delve deeper into discussing the findings of this research, what we can learn about our real-world test-cases from these findings, as well as potential explanations for certain findings. We suggest potential improvements for future research, and discuss how this study adds to the literature on Twitter-based public opinion research.

5.7.1 Why are New Hampshire models more accurate?

The most striking takeaway from our findings is the high model performance for New Hampshire versus the other two states. We suggest that there are four possible explanations for why this is the case: data quality, external factors, sampling, and Bernie Sanders.

First off, we suggest that the quality of data used in these analyses was simply highest for New Hampshire. We believe this to be the case as the New Hampshire presidential primary (of either party) occupies a uniquely significant space in the US political landscape. The fact that the first primary of every presidential election occurs here, results in pronounced media and political attention. Campaigns are made and broken in the state, and the media constructs new horse race narratives on a daily basis. This means that the race also receives a disproportionate amount of attention on Twitter, and we suggest that that increased awareness and resulting participation is likely to result in well-calibrated data for Twitter-based public opinion research.

The same story may also be shaping the model results by way of contributing externally to the samples from the other states. Given the states' primacy in the political sphere, and the close temporal proximity of the three states' election dates, much of the discussion on Twitter produced by users from South Carolina and Massachusetts in fact dealt with Bernie Sanders' victory in New Hampshire, rather than his or Clinton's actions or statements relevant to the primary in their state. While such statements are likely to indicate a given Massachusetts or South Carolina-based twitter user's support of Bernie Sanders in their own state, it isn't really a tweet about that particular election. This is highly difficult to disentangle in this particular research design. This factor also comes into play when considering sampling. As the time frames we selected for sampling means that all three states' sampling frames overlap, there is likely a lot of overlap between the three samples used in this paper's analysis. While we would ideally only sample tweets which are concerned with the given target state-level election, we can actually deduce from this that national-level news and events shape election-relevant conversation on Twitter, even if those elections are at the state-level.

In order to alleviate these sampling issues, there are a range of options for future research - first, a sampling frame is bracketed to be closer to a given election. While this may reduce the number of users who are included in a given analysis, it certainly increases the likelihood of tweets being relevant. Second, it may be fruitful to use machine learning classifiers to establish whether a given tweet is concerned with a given election or not. While this would involve further costly hand-annotation of tweets, it would likely to boost the power of any resulting model, as tweets could be selected with more confidence on their relevance to a given election.

Finally, there is the possibility that our New Hampshire models performed well because Bernie Sanders happened to win this state, and Bernie Sanders also happened to beat Hillary Clinton on Twitter, regardless of which state people were tweeting about him from. We revisit this point later, but suffice it to say that this clearly plays a role in this analysis, and says something about how Twitter is not a carbon copy of offline politics. However, it is also important to note that Sanders is far from a generic Democratic candidate, and simply because he dominates Twitter does not mean that any election would have one candidate that outshines all others on social media.

5.7.2 The impact of data processing and model specifications

Given the fact that this article employed both novel weighting and vote share percentage modelling strategies, it is useful to discuss their impact on the findings of this analysis.

The weighting mechanisms employed in this study can be understood to increase model performance overall. This is a greatly encouraging finding and supports the case for further work on estimating Twitter user-level characteristics for social science research. Overall, we find that, especially for location weighting, a theoretically plausible and consistent idea improves findings when transposing user-level political sentiment to aggregate measures of offline public opinion. However, especially in the case of South Carolina, we learn the limits of such an approach. This state, which is ranked 46th in the US for GDP per capita, has one of the lowest internet penetration rates in the country and is home to many older black citizens (who are among the least likely to be on Twitter), clearly isn't suitable for this kind of public opinion modelling, as the vast majority of opinions which may exist in the state simply never make their way onto Twitter. So, we can do the best possible job in adjusting samples, but if our sampling frame simply cannot include a large proportion of our target population, we will always fall short. Nonetheless, this is a clear indication that this kind of research needs to take sampling much more seriously, and consider "recruiting" users which satisfy certain criteria in a fashion akin to e.g. YouGov in order to improve models' predictive accuracy. But, most importantly, the South Carolina findings show a clear ceiling for Twitter-based public opinion research, barring a massive shift in Twitter membership in previously under-represented groups.

Next, it is important to consider how our model specifications and other design decisions may have affected the findings. First, there is the question of how the relative imprecision of our Ordered Logistic regression regressor may have exacerbated the relatively low quality of our data. We showed two core caveats with the regressors' accuracies: First, the best models had a mean absolute error of just below 2. This means that a) some positive tweets will have been labelled as negative and vice versa, and b) that in simulation-based models, several simulated votes will have been mis-allocated. Second, we showed, especially in the second-round validation stage where we selected the Ordered Logistic Regression model, that the regressors performed differently for the different candidates, over-estimating Sanders' average sentiment, while under-estimating Clinton's. Both factors will likely have contributed to the sub-par model predictions in the positive sentiment volume and simulation (full set) models. In order to overcome this in future applications, our intuition is that the only way to train more accurate sentiment regressors would be to rely on much larger training sets. It seems likely that n=1000 - even when as carefully constructed as our sets - do not sufficiently capture a sufficient proportion of the relevant features.

Furthermore, it is apparent that, at the row-level in Table 5, model predictions are broadly similar across states, despite the fact that they were computed using different source data. As touched upon earlier, we suggest this originates from the issue of what people are actually talking about in their tweets. It seems clear that users are talking about much the same across states, and indeed the underlying quantities - i.e. distributions of sentiment - are quite similar across states. While this may be adjustable with more intricate sampling (e.g. using the *a priori* recruitment of a representative set of users), and certainly the construction of larger samples, it may be a feature of this kind of Twitter data which is difficult, if not impossible, to alleviate.

Finally, we turn to the performance resulting from our novel method of transposing tweet-level sentiment to offline public opinion. It appears that withholding negative tweet scores from an analysis (i.e. the approach taken in existing research) works best when trying to get the best out of flawed data, in its optimal configuration. This is the case for both the South Carolina and Massachusetts models - we optimise the data we have, and can trust (seeing as we are only working with hand-labelled data) and get comparatively good results. The same is less apparent for the same method applied on machine-labelled data for those states, which is likely due to the low-quality data resulting from the overly greedy sentiment regressor. Conversely, for New Hampshire - which we believe is the best data we have in this sample, - we find that the simulation method (when only using trusted, hand-labelled data) matches the highest-performing model. So, while this is merely an intuitive conjecture and will have to be validated by future research, we suggest that the simulation method can add value to Twitter-based public opinion research when we have a trusted sample to work with, whereas the positive sentiment volume method is better at alleviating shortcomings in the data. Alternatively, there may be better ways of incorporating negative tweet-level sentiment into such analyses.

5.7.3 Lessons to learn about the elections from these analyses

There is a big elephant in the room when viewing the findings in Table 5: the candidates that participated in the elections, - in hindsight - enjoyed very different levels of popularity on Twitter, but also in the electorate, perhaps even in spite of their expressed political positions. When viewing these data, both in their raw form in Figure 4 or their aggregated form in Table 5, Bernie Sanders is clearly shown to be a highly appealing politician to the archetypal Twitter user who tweets about politics. Given what we know about the composition of the the platform's user-base and the consistent, left-wing track record of Sanders, this should not be particularly surprising. However, it is also highly likely that Sanders' disproportionate appeal, versus, for example a generic Democratic party candidate distorts this analysis. His political and campaigning style encourages participation on Twitter, and has been said to alienate those less enamoured with it. While this analysis clearly shows how a candidate like Bernie Sanders can use a platform like Twitter to their advantage, it is nonetheless important to keep in mind that he did not end up the Democratic presidential nominee in 2016, regardless of his overwhelming support online. More than anything, it would be interesting to see an analysis similar to this one replicated with other, more "conventional" and less divisive candidates.

Furthermore, this analysis makes it abundantly clear that a lot of people did not like Hillary Clinton at all. While the amount of tweets addressed to her were slightly fewer than those addressed to Sanders, the amount of negativity she received paints an overwhelmingly clear picture, which could have potentially alerted Democratic party elites to her shortcomings as a candidate before her defeat against Donald Trump in the 2016 general election. We suggest that this is true regardless of the fact that Twitter is not a representative sample, as its scale and geographic consistency is striking. This suggests that such types of public opinion research using Twitter data, even if they may not always paint an accurate picture of its offline counterpart, can be highly useful in better understanding a politician or candidate's appeals and shortcomings.

5.7.4 Contributions to the field and future research agenda

In conclusion, we outline this paper's contributions to the field of Twitter-based public opinion research specifically, and more broadly computational social science using digital trace data and highlight potential avenues for future research building on the findings of this project.

Perhaps our most encouraging finding is that weighting by user-level characteristics can help improve Twitter-based public opinion research. Especially for the "hard" attribute, user-level location, we find that models do better more-or-less across the board. The "soft" attribute, estimated user-level political affinity, also improves predictions in most cases. We suggest that future research should avail itself of these tools. Furthermore, we believe this paper makes a strong case for future research into computational methods for estimating latent user-level characteristics such as age, gender, ethnicity or educational attainment.

Less encouragingly, but no less importantly, we find that weighting cannot counteract under- and non-representativeness in source data. We suspect that a large proportion of older, black voters who ultimately gave Clinton her landslide in South Carolina had never used Twitter. While there may be mathematical methods to impute how those who are not represented will vote, it is a huge problem which is clearly illustrated by these findings, for which we have no obvious answer. However, we believe this to be a more nuanced appraisal of the feasibility of doing public opinion research using Twitter data than that of previous research, which (paraphrasing) stated "it doesn't work, so leave it". We believe that it is important to learn more about the opportunities and limits of certain research and modelling strategies, and this paper certainly proves a significant contribution in that regard.

This research also shows that simply sampling from a keyword-derived collection of tweets is likely not enough for public opinion-relevant insights which are to be valuable beyond Twitter. We argue that in order to build reliable models of public opinion from tweets, sampling has to be significantly more involved, considered and targeted. A useful avenue for research may be snowball sampling from established samples of users in a given geographic region, in a way that is agnostic to whether those users tweet about politics or not. We suggest that going from users to tweets is likely more fruitful than going from tweets to users.

A core improvement on future iterations of this research may be to enhance the machine learning component. We suggest doing this by classifying categorical sentiment variables at the tweet level. This would likely have a higher precision than the method used in this paper, and it could be useful to combine it with the predicted ordinal sentiment scores in order to better understand when the ordinal sentiment regressor makes widely erroneous predictions.

In further improvements to the models employed in this paper, we could have also trained a model encompassing all the hand-labelled tweets, regardless of their state of origin. While this would have likely led to some of the state-level intricacies being less pronounced in the model, a bigger training set tends to result in a better model. Furthermore, it seems fairly clear that there were not many state-level intricacies at play, because otherwise we would assume that the distributions of sentiment scores (Figure 4) would look different between states.

Chapter 6

Listening in on the noise: Estimating individual-level political preferences from tweets

Abstract

Twitter is used by people around the world to publicise their thoughts on politics. As the platform allows access to user-generated data, it is an increasingly important data source for political scientists, particularly those interested in public opinion. Previous research has used tweets to forecast election results, with inconsistent and unreliable findings which have usually been explained by the unrepresentative nature of the Twitter population, among other reasons. This has lead some to conclude that public opinion measured on Twitter cannot consistently be translated to the offline world. In this paper, I contribute to this debate by introducing a new method for estimating public opinion from tweets by shifting the unit of analysis from the tweet to the user-level. Specifically, users' voting intentions are estimated by training machine learning classifiers to predict voting intention as a function of how they discuss politically salient, election-relevant topics using distant supervision. This is based on the assumption that users' underlying "political views" shape both their political tweeting and offline political participation. Voting intention can thus be reverse-engineered by extrapolating from a small sub-sample of users who clearly declare their voting intention. I empirically test this method on the 2018 US midterm elections.

I apply this method on two distinct samples of twitter users (likely voters and random Twitter users), which forms a contrast to previous research, which typically relies on keyword-filtered samples of Twitter's stream.

When aggregated to the national and state-level, the best-performing classification model accurately predicts party-level popular vote share percentages with an error of 0.15%, thus exceeding the typical pre-election poll's accuracy. Furthermore, this research shows that a randomly selected sample of Twitter users produces the most accurate aggregations of individual-level vote choice estimates, providing a strong baseline for improved representativeness.

Keywords: Twitter, elections, public opinion, American politics, machine learning, distant supervision

6.1 Introduction

Since its launch in 2006, Twitter has established itself as one of the pre-eminent social media platforms, and a hub for many-to-many discourse about politics. It is the pre-eminent such provider that grants public access to content published to the platform by its users, which allows researchers to retrieve vast amounts of tweets and associated metadata, fuelling the widespread adaptation of Twitter data as the default data source for computational social science research using digital trace data (see e.g. Jungherr, 2015).

Twitter is used by individuals to post on most any conceivable topic, at a global scale. However, for political scientists, the most relevant (conceptual) domain of interest is *political Twitter*, where users talk about anything related to politics. Given the fact that every single day, millions of such 'political' tweets are shared, it is a plausible assumption that they can be aggregated, processed, and analysed in a way which enables researchers to extract an accurate reflection of public opinion, or at the very least, individual-level preferences from them. However, this is contingent on a thorough understanding of who it is that is sending these messages, and how well they represent a given target, offline population when aggregated. Overall, this endeavour - furthering the understanding of the utility of political tweets in isolation, within the context of their authors' preferences, within a context of similar users on the platform, and aggregated to a larger population forms the core contribution of this paper.

A useful test-case for any method aimed at estimating public opinion from Twitter data is the forecasting of election results. This scenario allows for the evaluation of the predictive accuracy of results, both compared to opinion polling, statistical election forecasts aggregating polls and fundamentals, and the election's eventual outcome. Indeed, given the fact that elections provide a ground truth quantity which aims to be closely approximated by forecasting, it is a useful test for gauging the general applicability of Twitter data for social science research designs seeking to measure, study or predict public opinion. In other words, if there is no way of making inferences from tweets in an area which is widely studied and inherently benchmarkable, they may not be useful for research in other areas of (computational) public opinion research. Furthermore, the relevance of this pursuit is emphasised by recent broadly publicised shortcomings and failings of traditional opinion polls and poll-based forecasting models, as well as widely observed declining response rates in surveys (see e.g. Groves, 2011). This brings under scrutiny the continuing reliability of survey-based methods.

The problem of generating and transposing measures of public opinion from Twitter to the offline world has received a large amount of scholarly attention. Starting with Tumasjan et al. (2010), several papers have attempted to forecast election results using Twitter data, with varying results. Existing research uses corpora of tweets collected with election-relevant keywords. Then, the share of tweets containing keywords associated with each candidate or party running in the election are aggregated, and in some cases mined for sentiment, excluding cases where relevant mentions are classified as *negative*. The proportion of (positive) mentions observed for a given candidate or party is then treated as analogous to their estimated vote shares in the target election, the main output of interest for forecasting models.

However, this tweet-level approach to forecasting election results fails to deliver consistent, reliable and theoretically grounded results. If Twitter, and political Twitter in particular, were online 'carbon copies' of their corresponding offline worlds, the tweet-mention-volume approach to extracting public opinion from tweet corpora would likely produce accurate and predictive election forecasts. However, they are most certainly *not*. Not only is the Twitter population distinctly unrepresentative of the voting-age population in several characteristics, such as age, gender, race, education, home location or partisan lean (e.g. Sloan et al., 2015; Smith and Anderson, 2018), there is also no guarantee that any given user will share their opinion during a given time frame of interest - indeed, some may repeat sharing their opinion several times a day, while others share nothing. These platform effects provide researchers with skewed datasets, where some bias can be known and corrected for, such as extremely active, often strongly partisan tweeters, while some bias cannot, such as most user-level socio-demographics, or an individuals' propensity to vote. Furthermore, even if we develop solutions for making adjustments along all the necessary dimensions, one crucial, and even more opaque uncertainty remains. Given the fact that data collected using Twitter's Streaming API[1] is likely to be down-sampled, and this down-sampling procedure is a black box to end-users of the API - there is no way of knowing how representative a tweet sample is of the whole population of relevant tweets (Morstatter et al., 2013). Jungherr (2017) refers to this as the "n=all fallacy" of working with digital trace data.

Following this, I argue that there is a need for a different conceptual, theoretical, analytical and empirical framework for using Twitter data for reproducible public opinion research. In essence, this involves a re-framing of the core research objective, away from a data-driven pursuit of forecasting election results using tweets, towards pursuing the more proximate goal of *measuring political preferences at Twitter's individual user-level*. While this research agenda contributes to the literature on Twitter-based election forecasts, it can be understood as more foundational in its focus: if tweets (and similar user-level digital trace data) *are* a primary data source for political science research, and more generally computational social science research concerned with individual- and group-level behaviour, then it is of utmost importance to understand *what tweets signify* and hence how they can be used to understand societal and political phenomena of interest. While this in itself does not address the issue of representativeness, I argue that it is a necessary condition for the production of accurate insights from Twitter data, be the samples representative or not.

In this paper, I approach this task from the perspective of public opinion, measured at the individual level, in the context of elections. While little consideration has been given to this in the existing Twitter-based forecasting literature, it is clearly a necessary, while not sufficient, condition for producing accurate, and more importantly, reproducible election forecasts, as, ultimately, it is *people* who vote - not tweets. For this purpose, this research follows an inductive framework: data are processed and analysed in line with known, empirical quantities, and then fit to a known, true outcome quantity to evaluate the efficacy of resulting predictions.

This paper's core contributions are threefold, and mostly of a methodological nature. First, I introduce a framework for extracting public opinion from tweets by shifting the frame of analysis from the tweet-level to the user-level. Users' tweets are collected over a pre-defined time period leading up to the date of a target election. Then, tweets which are likely to denote users' self-declared voting intention are extracted from the corpus of all

[1] API = Application Programming Interface, i.e. a means for developers to access data from other applications through the web. See https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/introduction (accessed 11/01/21) for more on Twitter's Streaming API

collected tweets, and hand-annotated to reflect the users' declared vote choice preference. Second, these self-declared, individual-level voting intention variables are used as training data in a 'distant supervision' machine learning classification algorithm. In other words, the algorithm is able to probabilistically classify user-level voting preferences for users who *did not* explicitly state it in their published tweets. This is achieved by filtering all sampled users' available tweet text for relevant, politically salient, pre-defined *related* topics, which were extracted by matching tweet text to pre-defined keywords, for example, the names of party leaders, the names of parties, or terms which are used to describe the most salient election-relevant events during the time leading up to the election. Most importantly, the topics have to be political in nature and related to the area in which public opinion is to be measured - in this case, the target election. The core assumption of this method is that individuals who prefer different parties or candidates will also differ in the way they express their opinion regarding related topics. Hence, this research design is a process of reverse-engineering users' published, politically salient content as an instrument for computing electoral preferences. Thirdly, this paper adds significant evidence regarding the impact and importance of sampling decisions in Twitter-based public opinion research. The importance of sampling is under-studied in the existing literature, to the point where it is often all-but ignored, which in turn hinders progress in the area of furthering the understanding of Twitter's (un)representativeness. Here, I compare identically applied data processing on two distinct samples: one purposive, selected on having tweeted something indicating a users' participation in the 2016 election (i.e. targeted selection of previous voters who are likely to vote in the future), and one random, where user-ids are selected using a random number generator. Both samples are reduced to users from the USA using my geo-locating pipeline (Loynes et al., forthcoming). While this selection of disparate datasets by no means alleviates the issues of representativeness associated with Twitter data, it seeks to further the understanding of how advanced sampling approaches can address these issues.

These three-fold methodological innovations are applied on the test case of the United States midterm elections of 2018. Midterm elections are useful for this pursuit, as they allow for a geographically dis-aggregated analysis of findings, while keeping the fundamentals of the election - electoral rules, competing parties, election date - broadly constant. Indeed, the different underlying state-level specifics can illuminate the relative salience of different, issue-based causal factors contributing to aggregated public opinion in that geographic entity, and further provide contextual evidence to evaluate a method's accuracy. However, the midterms may not be the ideal test case for producing an 'easy' study of Twitter-derived public opinion, as turnout is typically lower and more volatile than in, e.g., presidential elections, and there is no clear consensus in the established political science literature regarding the key factors shaping the elections' eventual outcomes, besides a prevailing theme of voters often 'punishing' the incumbent president (see e.g. Campbell, 1991). However, as previously stated, this paper's objective is not to predict the outcome of the midterms for the sake of doing so. Rather, it is to introduce a reproducible method with which any Twitter user's likeliest preference in a given politically and public-opinion-salient issue-area can be estimated. For this paper, the dependent variable is individual vote choice - but the variable to be measured could be anything typically aggregated into public opinion, such as party identification or institutional trust. The only requirement is that a target user has tweeted relevant, politically salient content which can be extracted with topic-level keywords for distant supervision.

In the following sections, I introduce the existing literature on Twitter-based public opinion research and election forecasting, focusing on their relative efficacy. I then summarise the key takeaways for this project from the existing literature and further outline how my proposed framework addresses them. I then provide a detailed description of my proposed method for estimating individual-level political preferences as a function of users' tweets. After describing the sampling method used to generate the samples used in this paper and providing descriptive statistics on available characteristics of these samples, I present and discuss my research findings in regard to model and classification performance, and further present aggregations of estimated user-level vote choices to the national and sub-national levels outlining its usefulness for the endeavour of estimating individual-level political preferences.

6.2 Related Work

Since its inception in 2006, Twitter's promise as a global, large-scale source of public opinion data has resulted in numerous political science papers on the subject. Overall, this literature can be divided into four distinct sub-groups:

1. Papers describing approaches to eliciting public opinion from Twitter (typically framed as as an alternative to surveys)
2. Papers documenting the use of Twitter data to forecast (or post-cast) election results
3. Research introducing methods for estimating latent user-level characteristics, such as race, gender, political ideology or location
4. Papers which outline the uses of Twitter as a tool for influencing public opinion

Below, I summarise key findings from the respective sub-categories. The focus is on research that utilises Twitter data in some form for measuring public opinion, or situates its contribution in making such measurements more reliable or easier to obtain. For the purpose of best placing the contribution of this paper, the focus lies predominantly on Twitter-based election forecasts, as this category ties in most directly to the subject matter of this research, especially when regarding research methodology.

6.2.1 Twitter as a tool for measuring public opinion

The promise of data gathered from Twitter as a source of information on public opinion is intuitively plausible. The sheer number of information-rich data points containing unsolicited, individual-level statements on most any topic offers the opportunity of understanding what target populations think and want without having to ask them. Evidence for this purported linkage is provided by O'Connor et al.'s (2010) influential study on the correlation between long-term tweet sentiment[2] on US president Barack Obama, and the analogous 'offline' measure, presidential job approval ratings. The authors find a significant correlation of r=.77

[2] Sentiment analysis is a linguistic method whereby text units are mined for their emotive/affective sentiment toward the text's object. This can be undertaken with different methods, most broadly delineated into lexicon-based approaches, which match text units to lexica of words pre-annotated with their respective sentiment polarity and intensity, or by training machine learning classifiers on sub-samples of target datasets to be annotated and classifying the entire sample thusly.

to r=.81 (p. 128), supporting the hypothesis that data derived from tweets can be used to measure public opinion.

Several authors have expanded upon O'Connor et al.'s (2010) findings, with a diverse array of substantive applications and methodological approaches. For instance, Kim and Kim (2014) analyse Korean-language tweets with the goal of parsing South Koreans' views on military tensions with North Korea at the time. The authors argue for Twitter data as a supplementary source to polling data, and especially one which, through its fertile landscape for the spreading of rumours pertinent to dominant political issues, can illuminate individuals' perceptions and opinions on a crisis, where opinion polling may choose not to address such issues. Cody et al. (2015) and Cody et al. (2016) conceive of Twitter as an "unsolicited public opinion poll" (2016, p. 1), with the power of both mirroring measurements in traditional polling, but also gleaning insights into matters which are rarely surveyed. In a similar methodology to O'Connor et al.'s (2010), the authors find a strong correlation between long-term sentiment on twitter and corresponding opinion polling. Mejova et al.'s (2015) edited volume "Twitter: a digital socioscope" assesses the state of the field of Twitter-based public opinion research and highlights applications in different issue areas - political public opinion, economic sentiment, public health, psychological indicators and disaster detection/response. The measurement of economic sentiment in the public (not to be confused with tweet text sentiment analysis) through Twitter data has further achieved broad scholarly attention, e.g. as a way of tracing and predicting stock market developments (Bollen et al., 2011b), or to measure the public's evaluation of US presidential candidates' economic platforms (Karami et al., 2018).

The existing literature on Twitter as a means for measuring public opinion consistently shows strong correlations between Twitter-derived measures of public opinion on a broad array of diverse issues with analogous offline measures. This provides a strong case for the usefulness of Twitter data as a source for public opinion measures. However, this research typically approaches these topics from a high-level, 'bird's-eye-view', meaning that huge amounts of tweets are analysed over an exceedingly long time-span, and mined for sentiment at the aggregate level. Further, areas of interest within public opinion are framed as yes-no / approve-disapprove issues - which neatly allows for comparisons between positive-negative sentiment-classified, year-long tweet text and opinion polling data. However, I suggest that these approaches, while useful in providing foundational evidence for the field at large, would prove less useful in an applied context.

6.2.2 Twitter-based election forecasting

The issue area within the field of Twitter-based public opinion research which has arguably garnered most attention is focused on using data derived from Twitter to predict/forecast some quantity delineating the result of a given election. This area of the literature builds, sometimes explicitly, often implicitly, on the entire scope of public opinion research with Twitter data. While my research in particular ties in thematically more closely to the general Twitter-based public opinion literature, it is methodologically and in terms of output most closely related to what has been attempted here.

While studies forecast (or post-cast) different elections, the papers can be divided into a small number of distinctive sub-groups based on the methodological approach applied: *volume-based forecasts*, whereby the relative volume of the occurrence of keywords in tweets associated with competing parties/candidates is used as a proxy for eventual vote share; *sentiment-based forecasts*, where tweet text featuring relevant keywords is first mined for its sentiment (positive, neutral or negative) towards the keyword (e.g. candidate), and proportions of *positive* tweets are used as proxies for vote shares; as well as other approaches, which employ various supplemental data sources (e.g. Google Trends data or polling data) in order to produce a forecast, or measure vote shares by analysing follower numbers for candidates'/parties' verified social media accounts.

Measuring the ***volume of keyword frequency*** in tweet collections is arguably the simplest way of producing a Twitter-based election forecast. Whichever keyword (or set of keywords) associated with a given candidate or party occurs most frequently is understood as signalling the candidate/party with the highest vote share. The first (and widely cited) paper employing this approach covered the German Bundestag elections of 2009 (Tumasjan et al., 2010). The authors found that relative mention volume, using party names and party leader names as keywords, predicted the eventual vote shares with a mean absolute error (MAE) of 1.65%, in line with traditional opinion polling. However, Jungherr et al. (2012) highlighted the shortcomings of such as design: If the Pirate Party, which received a vote share of approximately 2% in 2009[3], had been included in Tumasjan et al.'s original analysis, this party would have been the winner and largest party. This highlights one of the key shortcomings of volume-based approaches: *Researchers' design decisions and inclusion criteria have a significant influence on the accuracy of the forecast*. In retrospect however, it is arguable whether a similar design decision would produce similar results if it were repeated today. In the following years and elections, the Pirate Party has become widely obscure in Germany while Twitter use in the country (and globally) has risen significantly. Perhaps, then, the unique case of the Pirate Party being home to and loved by early adopters who were also hugely influential on German-language Twitter at the time, is something that would not repeat itself nowadays, and would not translate to different party systems. At the very least, however, it illustrates a huge complexity of using Twitter as a data source - the platform evolves dynamically, beyond not only the control of its managers or individual users, but also beyond their timely comprehension of change. To assume that anything which was 'true' mere years ago can be translated to the current context without considering all available factors influencing temporal decay of research findings drawn from a different state of the platform from which data originate is likely incorrect (see e.g. Munger, 2018)

Other scholars have used similar volume-based designs, however aiming to avoid similar errors by more rigorously specifying keywords, or only tracking hashtags known to be associated with the election or relevant candidates/parties (e.g. Caldarelli et al., 2014; Cunha et al., 2014; Jungherr, 2013, 2014; Nooralahzadeh et al., 2013), while other authors included volume-only forecasts in their publications in order to compare them with other approaches (e.g. Sang and Bos, 2012). A particularly insightful piece of research in this vein is DiGrazia et al. (2013), where the authors use Twitter mention and user-numbers for US Congressional candidates in 2010 and 2012 in regression models predicting vote share in those elections. They find that, even in models controlling for % of Republican vote in district, median age, ethnic composition of district, % college educated in district, median household income in district, and a news consumption variable for the district, that both the raw volume

[3] and, due to the German constitutional arrangement of a necessary 5%-vote share threshold which needs to be reached in order for a party to enter parliament, did not end up sending any MdBs to the Bundestag

of candidate mentions and the number of users mentioning a candidate have a highly significant positive impact on candidates' vote share, across all 795 elections studied. This supports the intuition that political tweets do contain signals relevant to understanding offline public opinion, regardless of the unrepresentative nature of the platform.

The most accurate applications of the volume-based method achieved mean absolute errors of under 2%, which is in line with traditional pre-election opinion polling (e.g. Tumasjan et al., 2010; Sang and Bos, 2012). However, overall, this approach is unstable and unpredictable in the sense that the conceptual framework of the linkage between observed tweet text and vote choice is a black box (Loynes and Elliot, 2021), and accuracy is very much influenced by seemingly arbitrary design decisions taken by the researchers. As Ceron et al. (2016) put it in their comprehensive book on the state of the field: "no matter whether they record attention, awareness or support, merely computational data seem to retain some problematic attributes and might fail to catch the informational complexity of the social media environment." (p.20). This highlights an interesting question which has been investigated empirically (see e.g. Jungherr, 2014), but not answered comprehensively: what exactly does keyword-derived volume tweet data signify? Most likely, it is a combination of all three (attention, awareness, support), but other than support it can also contain its inverse: dislike or disapproval.

To summarise, volume-based methods offer a tantalisingly simple means of producing Twitter-based election forecasts, but they do not consistently work. However, volume data *may* have a useful application in Twitter-based public opinion research, when they are understood as a measure of attention/awareness (pure volume) or support / dislike (positive / negative sentiment volume). Furthermore, DiGrazia et al.'s (2013) work shows a consistent pattern of 'more tweets, more votes'. In other words - party or candidate-level tweet mention volume should not be ignored when studying Twitter-derived public opinion, it is just not telling the 'whole story'.

Sentiment-based forecasting approaches

If large-n collections of tweets concerning political candidates in elections are indeed indicators of "attention, awareness or support" (Ceron et al., 2016, p. 20), then the goal of incorporating sentiment analysis into election forecasts can be understood as a filter that isolates the "support" component of such data. O'Connor et al. (2010) found that tweet sentiment correlates highly, ranging from r=.77 to r=.81 (p. 128), with opinion poll time series on presidential job approval, adding strong evidence to the positive-sentiment=support hypothesis. Several publications have since sought to forecast election results using tweet volume data mined for positive sentiment. Bermingham and Smeaton (2011) forecast the Irish general election of 2010 using supervised sentiment analysis, as well as benchmarking their analysis with mention data, but predicting parties' vote shares by adding all variables into regression models (fitted to the polls). The best MAE achieved by these models does not match that of traditional opinion polling, always exceeding 3%. Sang and Bos (2012) follow a similar paradigm with the Dutch Senate elections, whereas their measure of accuracy is at the Senate-Seat-level, where 8 seats out of 75 (p. 59) were falsely assigned to a given party. Furthermore, they stratify their sample by only allowing one tweet per tweeter, and the mention of only one party per tweet.

The most methodologically rigorous (and predictively accurate) research in the sentiment-based forecasting domain to date was conducted by Andrea Ceron and colleagues (Ceron et al., 2014, 2015). The authors use a supervised sentiment analysis algorithm based on the work of Hopkins and King (2010). Using this approach, the authors were able to forecast the outcome of the 2012 US presidential election with a MAE of 0.02 % (Ceron et al., 2015, p. 11). Furthermore, they were able to forecast vote shares in crucial swing states more accurately than the average of pre-election opinion polls in 8 out of 12 cases. They achieved these results by filtering tweet collections to only include statements containing a clearly extractable voting intention as well as a *positive* party or candidate mention. Loynes and Elliot (2021) provide an in-depth comparison of different applications of sentiment-labelled data for estimating vote share percentages in elections, whereby their most predictive models achieve MAEs below 1.

On aggregate, sentiment-based methods have higher predictive accuracy than volume-based methods. However, there is no one perfect sentiment analysis tool for measuring sentiment in relevant tweets. Indeed, there is no clear consensus as to what sentiment analysis for tweets is actually measuring - is it a linguistic classification problem seeking to analyse the intensity and polarity of *language* without analysing what such language signifies in a given context, such as politics or elections? Or is the classification task seeking to extract "political sentiment" or *stance*, in other words what language means in the specifically political domain?

Other approaches

Some researchers have sought to leverage data other than that derived from the content of social media postings. For instance, Barclay et al. (2015) found that the volume of "Likes" on verified candidates' Facebook pages were a strong predictor of their vote share in the 2014 Indian Lok Sabha elections. However, this finding is not in line with Giglietto's (2012), who forecast the Italian mayoral elections of 2011 with the same methodology - here, no significant correlation with Facebook likes and vote shares was observed. Ceron et al. (2016) note that candidates' follower numbers - the Twitter equivalent of Facebook page likes - should not be understood as indicative of vote shares, but rather of attention to a certain candidate (p. 19). They use the example of Barack Obama and Mitt Romney's follower numbers on Twitter during the 2012 US presidential election: Romney was out-followed by Obama by a factor of 17, while the election result was considerably closer. Lui et al. (2011) found that Google Trends search volume data was not a good predictor of vote shares in the 2008 and 2010 US elections, suggesting that these data fall short of Twitter volume data in their predictive accuracy of elections. Franch (2013) combined data from Facebook, YouTube, Google Trends and Twitter with polling data to produce very accurate vote share forecasts using multi-level auto-regressive models. Beauchamp (2017) used features extracted from Twitter volume data, subset by individual tweets' authors' US state locations, to predict and interpolate state-level polling leading up to the 2012 US Presidential election. The author found that Twitter mention data can indeed be successfully used to predict and interpolate polling data, but the predictive accuracy declines the further away the time frame of tweets is from the target poll. Wang et al. (2015) used data obtained from daily vote-intention polls given to users of the Xbox Live gaming platform to show that when statistically processed in the right way, explicitly non-representative public opinion data can produce election forecasts on par with those derived from representative polls.

6.2.3 Estimating latent Twitter user-level characteristics

While there is an aggregate-level understanding of the demographic composition of social media platforms (see e.g. Center, 2018), this is informed by survey-based findings, and is thus not easily useable for translating insights gleaned from Twitter data onto some mapping of public opinion. Hence, in order for public opinion research based on Twitter data to be reliable and useful beyond the birds-eye-view approach outlined above, there is a need for the estimation of individual-level (socio-) demographic attributes. This was first highlighted by Mislove et al. (2011), who developed approaches to measuring the makeup of Twitter's population regarding three relevant attributes: users' home locations, their race/ethnicity and their gender, by matching user-level metadata to US census records. Crucially, the authors find that the population of Twitter users is distinctly different from the offline US population in that it skews more to younger, well-educated, liberal, urban and white people. Since this groundbreaking research, several studies have introduced new methods for inferring user-level characteristics. When relating this to public opinion research, it is important to highlight Pablo Barbera's work. Barberá (2015) outlines a highly influential method for estimating Twitter users' *political affinity* on a two-dimensional (left-right) axis using the assumption of homophily in humans' social ties. Further research in this vein has been conducted by Bond and Messing (2015), who use endorsement statements of politicians on Facebook to build models by which they are able to propagate ideology estimates to both active US politicians and 6 million users. This measure correlates highly with both established measures of US politicians' ideologies and users' self-reported ideology-placement on Facebook. An alternative method for estimating use, politician *and* URL-domain-level ideology is introduced by Eady et al. (2019), who use an item-response approach to inferring latent ideology from individual-level URL-sharing behaviour on Twitter.

Besides his work on ideology, Barberá (2016) introduces methods for matching user-level information from Twitter with publicly accessible voter records, allowing for the production of large-scale samples with rich demographic information attached at the user-level. However, the sparse availability of public voter registration files, even within the context of US states, means that this method is not generally transferable to a broad range of contexts and applications. Besides Barbera's work in this area, there exist a number of studies which match data taken from Twitter to an existing, census-derived offline data source. For instance, Mullen (2018) present a method (complete with open-source R package) for probabilistically estimating US Twitter users' gender/sex by matching their first name (as reported in their profile data) to US Census records. Similarly, Imai and Khanna (2016) provide a method for US users' ethnicity, however this time using last names. As both methods provide probability estimates for each ethnicity category, the user can decide their own cut-off point.

Several papers in this area use machine learning methods to classify probabilities of individual users having a given demographic attribute. For instance, Conover et al. (2011) predict user-level political ideology for a large sample of Twitter users using support vector machines, while Al Zamal et al. (2012) classify age, gender and race with a similar, machine learning approach. Volkova and Bachrach (2015) shows how machine-estimation techniques for a wide range of individual-level attributes, such as gender, age, religious affiliation or education level vary, and this variation begins at the data annotation stage, whereby certain attributes such as race/ethnicity achieve significantly higher inter-rater agreement than others, such as religious affiliation.

A large amount of attention in this research area has been devoted to the area of inferring users' home locations, again with varying approaches, from census-record-matching, to machine learning based on tweet and metadata content, to leveraging the Twitter network. Loynes et al. (forthcoming) synthesises the most accurate elements of the state of the art into a reproducible data pipeline that can be used on any given Twitter user, and results in granular location information with attached uncertainty estimates.

Finally, Hinds and Joinson's (2018) paper features a systematic review of research on demographic estimation through digital trace data (including, but not limited to Twitter and social media data) to date, and highlights the disparity between multiple approaches in regard to their accuracy. However, it also showcases how seemingly banal records and traces of humans' behaviour and interactions in the digital age can reveal a lot about users.

6.2.4 Twitter as a tool for *changing* public opinion?

How can Twitter (and similar services) be used to *change* public opinion, e.g. from the perspective of an office-seeking political candidate, or from a policy-seeking organisation or party? Building on early, groundbreaking work on the approaches taken and impact of 'cyber-campaigning' in the Australian electoral context (Gibson and McAllister, 2006), which finds that this style of engagement on the side of the candidate has an independent, measurable, positive impact on candidate success, many scholars have approached the question of the efficacy of online (Twitter-based) campaigning from different perspectives and with different research designs. For instance, both Grant et al. (2010) and Kruikemeier (2014) find that candidates who actively make use of Twitter to disseminate their message manage to increase their vote shares (found in contexts as different as Australia and the Netherlands). Vergeer et al. (2011) conduct a similar study, however this time in the context of the 2009 European elections. They find that, "the more frequently candidates tweeted, the more votes they received (r = .318, p <.01). [...] Candidates who increased their blogging closer to Election Day also received more votes (r = .307, p < .01)" (p. 16). At this point, however, it is important to note that this effect was not found in the US context, where Hong and Nadler (2011) conduct a similar test of the impact of tweets on candidates' vote share. While there is a legitimate question of whether these findings still hold true in the same way in 2021, it is nonetheless striking that these findings were replicated across such different electoral contexts. The fact that it wasn't replicated for the US may indicate the larger degree to which non-institutional campaign finance dominates the success of candidates in the US context. Indeed, I suggest that nowadays, the effect of tweeting as a campaigning tool is likely less pronounced, but the effect of *not* making use of it is likely measurably high. Finally, for a comprehensive review of the state of the art of digital campaigning (beyond merely the realm of Twitter), see Vaccari (2013).

6.2.5 Tentative conclusions

This literature review clearly shows that politically salient Twitter data contain signals pertaining to public opinion. Further, I have shown that there have been varied approaches to extracting such signals, as well as transposing them to the offline domain. This is the case when framing tweets as quasi-polls, when aggregating tweets to forecast elections.

However, the patchy track record of Twitter-based election forecasting indicates that a sole reliance on the basic metrics extricable from politically (electorally) salient tweets are not sufficient, further exemplified by the modelling presented in DiGrazia et al. (2013). In order to use them adequately in public opinion measurement or election forecasting settings, this means that either they have to be supplemented with external data, ***or signals have to be extracted from them differently***. In the following section I outline how I do this in this paper. However, it is useful to first understand two core challenges that must be addressed in order to add useful insights to the literature on Twitter-derived public opinion: First, signals need to be disentangled from the inherent noise resulting from *ad-hoc* language, but also from the more general noise that is the Twitter stream. In other words, the challenge lies not only in identifying the needle (pertinent tweet) in the haystack (twitter stream), but also to remove it from whatever it may have got caught in and to clean and polish the needle. The second challenge lies in adjusting samples of tweets (or public opinion-relevant signals extracted therefrom) to better reflect target populations, something that has received surprisingly little attention in the area of Twitter-based public opinion research. This is especially surprising given the considerable amount of work devoted to estimating Twitter user-level characteristics, a necessary condition for reliably adjusting samples.

While there is no guarantee that tweet-derived measures and aggregations of public opinion can be reliably and consistently produced, there remains ample scope for novel exploratory research into the nature of signals contained in political tweets, and their usefulness for public opinion research. Despite its mixed record in the past, I argue that the election forecasting paradigm is a fruitful avenue for achieving this. However, I argue for a range of enhancements which I implement in this paper:

1. Analyse data at the user-level, not the tweet-level (as ever, it is humans who vote, not tweets) - very few previous studies have done this, although those that did proved most accurate (Ceron et al., 2015) or insightful (DiGrazia et al., 2013).

2. Be mindful of the n=all fallacy (Jungherr, 2017), and hence use different sampling approaches to investigate the relative efficacy of a given approach. Simply using unadjusted large-n samples of keyword-filtered tweets derived from the streaming API is unlikely a good enough starting point, as there is no guarantee the sample accurately reflects the entirity of Twitter, let alone offline populations.

3. Expand the analysis beyond mention volume. How can support/opposition and attention be disentangled? Is there a quantity derivable from pertinent tweets with a higher correlation to the target quantity, ***votes***?

4. Incorporate measures of user-level (socio)-demographics into the analysis. There are several tools available, so there is no excuse not to use them. This is a necessary building block for conducting more representative and thus accurate Twitter-based public opinion research.

6.3 Research Design

6.3.1 Toward a framework for extracting individual-level political preferences from tweets

In this paper, I introduce a framework with solutions for both challenges outlined above - the identification of pertinent tweeets and the extraction of relevant signals as well as accurate sampling/sample adjustment of Twitter collections. First, I demonstrate that individual-level political preferences can be computed at a by-user basis by generating small samples of hand-labelled ground truth data on users' voting intentions where this is confidently observable, and using this as training data in machine learning models to predict the same for users where it is not. This is achieved by leveraging tweets with related, politically salient topics in machine learning models, whereby the hand-labelled voting intention data is used as a distant supervisor. This hypothesises that individuals vary in how they discuss politically salient issues as a function of their underlying, latent *political views*. In other words, the way a person views e.g. the necessity of government services and taxation will inform their political, and, crucially, electoral preferences. This method is scaleable and constrained only by the availability of politically salient, user-published content in a time-frame of interest.

Second, I outline the next step in this process, which involves a sampling framework allowing for these computed metrics to be aggregated so as to reflect target populations in the offline world. By randomly sampling large numbers of Twitter users and classifying them in a number of key quantities of interest, such as their sex/gender, ideology/partisanship or location, we can construct large 'civic panels' of Twitter users, which, in conjunction with the method of estimating individual-level political preferences described in this paper, will allow us to dynamically track public opinion on Twitter, and weight it to reflect target populations. In this paper, I pilot this sampling approach by creating multiple, distinct samples of Twitter users: one 'civic panel' of purposively sampled likely voters who are classified by their sex, ideology and location, and two, a 'control panel' of randomly selected Twitter users who were geo-located into the USA. This allows for a comparison of how differently composed samples vary in regard to the distribution of vote choice classifications. By investigating factors shaping inter-sample variance in computed public opinion estimates, this can further feed into future work employing this methodology for larger samples seeking to approximate target offline populations, and more generally illuminate the importance of sampling in the big data age, as it provides a way of mitigating the effects of the n=all fallacy. However, further work in developing and improving demographic estimation tools for Twitter users as well as understanding the composition of the Twitter population is needed in order to achieve the most representative samples possible.

6.3.2 Method

In this research, the key outcome variable of interest is the probability of an individual Twitter user voting for a given party (operationalised as Democrats or Republicans) or candidate in a given election (in this empirical application, the 2018 US midterms). This probability is estimated as a function of how users talk about related politically salient issues on Twitter. Besides party or candidate preferences, this method can be applied to extracting any given individual-level probability of being in support or opposition to a candidate, policy, or similar. The proposed method requires a sample of individual Twitter users, for whom all tweets authored in a pre-defined time-frame are collected. These data are then filtered for tweets concerning the concept that is to be studied (in this case, voting intention). The resulting sample (or, if it is too large, a sub-sample thereof) are then hand-annotated to reflect the value of the given concept to be measured, as deduced from the tweet text.

Defining the quantity of interest and extracting data for labelling

While most users do *not* state publicly whom they intend to vote for, some do, and in some cases they do so repeatedly and explicitly[4]. For the purpose of building a ground truth training set of users with clearly classifiable voting intentions, I extract these pertinent tweets by filtering all sampled tweets by relevant keywords[5]. This provides a high-precision approach for extracting relevant statements, but it is likely that relevant tweets are not captured due to being phrased differently[6]. These filtered tweets are then hand-coded for individual-level voting intention. Each possible outcome is to be labelled in a binary fashion, i.e. **X** or **not X** (denoted as !X). Hence, in a scenario with two parties, this results in the following labelling categories (with real-world examples in parentheses):

1. X (Democrats)
2. !X (*not* Democrats)
3. Y (Republicans)
4. !Y (*not* Republicans)
5. Abstention
6. !Abstention
7. not classifiable / 'don't know'
8. not applicable

X and Y are mutually exclusive, and, in a two-party system, !X typically implies Y and !Y typically implies X, but both can also mean abstention. Abstention, X and Y are mutually exclusive, but !Abstention means either X or Y, but is often equivalent with 'not classifiable'. This can signify any value, but the available text does not contain sufficient information to confidently assign it a substantively meaningful category. Furthermore, the categories X, Y and Abstention are substantively relevant, whereas !X, !Y, !Abstention can all also mean 'not applicable', meaning that it is not recommended to use these categories to train classifiers. Indeed, a tweet labelled 'not applicable' may contain content unrelated to the labelling task at hand, such as "In the teen choice awards, #Imvoting for Justin Bieber" - this tweet contains no information about the author's preference in an upcoming election.

[4] While there is no reason to believe individuals who share such information could be lying, or choose not to act on such an intentional statement come election day, this is a possibility. However, this is also the case for surveys: survey respondents have been shown to lie (e.g. Guess et al., 2018), and factors which may influence lying in surveys are not present in social media data

[5] The keywords are: *"my vote", "i'm voting", "i am voting", "i'll be voting", "i will be voting", "i am going to vote", "i will vote", "we'll vote", "we will vote", "we are voting", "i choose to vote", "im voting", "ill be voting", "myvote", "myvote2018", "imvoting".*

[6] Recall for extracting voting-relevant tweets from samples can likely be improved by developing machine learning classifiers which can detect statements of voting intention in tweet collections

Defining related topics and preparing data for analysis

While most Twitter users do not explicitly share who they intend to vote for, many do tweet about politics. The nature of such statements are shaped by authors' political views: a liberal talks about Donald Trump in a different manner to a conservative. I posit that these latent *political views* are a key factor explaining *both* how users discuss politics on Twitter *and* how they will likely vote in the election. This is visually depicted in Figure 1. By observing the variance in how individuals discuss relevant political topics on Twitter, we can then estimate their likeliest vote choice as a function of the language used when discussing politically salient related topics.

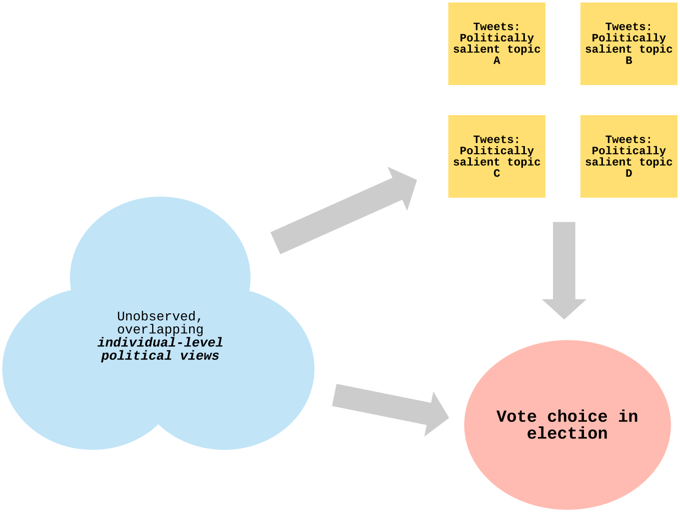

Tweets: Politically salient topic A
Tweets: Politically salient topic B
Tweets: Politically salient topic C
Tweets: Politically salient topic D
Unobserved, overlapping *individual-level political views*
Vote choice in election

Figure 6.1: Model of shared factors explaining both vote choice and political tweeting

It is important, then, to clearly define what such related topics may be, and why. Do partisans of opposite persuasions discuss the weather in systematically different ways? Perhaps, but if so, this is likely not due to their underlying political views, but rather due to their likelihood of living in different areas and thus being used to different weather. Accordingly, when identifying topics of interest influenced heavily by an individual's political views, the easiest targets are politically salient issues. Furthermore, it is useful to define issue areas which are transferable to multiple applications, both temporally and spatially. For this purpose, I define the following topics of interest which are employed in this paper in general, as well as their specific implementations in the context of the 2018 US midterms (in parentheses). This selection can be expanded or substituted, as long as the goal of widespread applicability is kept in mind.

1. ***Leading politicians.*** Typically, these would be the head of state or government, or party leaders. As they are the foremost ambassadors of their partisan brand, they are likely to be praised and endorsed by their supporters and criticised and maligned by their opponents. (For this paper, I chose Donald Trump as the key polarizing figure. Keywords: *"trump", "donald", "realdonaldjtrump", "potus"*)

2. ***Relevant news events.*** Most any electoral campaign will feature major news stories which have an influence on the election. Again, differently aligned partisans view and

interpret these events differently. (For this paper, I chose the controversy surrounding the nomination of Supreme Court Judge Brett Kavanaugh[7]. Keywords: *"kavanaugh", "brett", "supreme court", "scotus"*)

3. ***Parties and/or candidates.*** While this topic category may be subsumed by “Leading Politicians”, this is not necessarily the case and depends on the electoral rules and institutional configurations of the election of interest. The same hypothesised mechanism however holds true, in that partisans will talk about their “team” differently than they talk about their opponents. (For this paper, I chose to only filter by party names, as the number of candidates running for all offices in the midterm elections was exceedingly large and varied by state. Keywords: *"democrat*", "dems", "republican*", "gop", "reps"*)
4. ***The election.*** This topic is the most clearly related to the quantity of interest. However, it is not immediately clear that conversation about the election itself would differ significantly given different political views. However, this may also act as a useful reference category, and it is most certainly relevant. (For this paper, I chose keywords which clearly refer to the 2018 midterms. Keywords: *"midterms", "midterm", "2018 election", "gubernatorial election", "senate election", "house election"*)

All sampled tweets are filtered by the specified topic-level keywords. Again, note the importance of clearly conceptualised, theorised and specified keywords in order to be able to confidently interpret values computed with text featuring these keywords as input data. Noise and uncertainty will always be a factor when working with political tweets. However, the process of rigorous, extensive thought devoted to the concepts relevant to one’s dependent variable are an essential process in any research design in computational social science, and has, as I have argued previously, been somewhat neglected in much applied Twitter-based public opinion research.

It is important to decide on a time range from which to filter tweets for keywords: full sampled tweet collections spanning the entire timescale of interest or temporally specific subsets. While the former will provide larger amounts of data and thus predictive text features, the latter allows for more narrowly specified classifiers, and further allows for the analysis of individual-level covariates of topic-level tweets. In practice, this decision boils down to two considerations. First, if one is seeking to track change over time, the latter is likely best suited for this endeavour. This may also be necessary if the language used to discuss a given topic is expected to change over short periods of time, as may be the case when dealing with the dynamic nature of political discourse on Twitter surrounding an election. Second, this decision is significantly constrained by the availability of training data. If sufficient (training) data for shorter time increments are not available then temporally segmented classifications will not be reliable. For the purpose of this research, I choose to train and apply classifiers on all available data, from a period of two months leading up to the election date.

Training models

Having labelled relevant ground-truth data (to function as training labels in binary classification models), defined relevant politically salient topics of interest (which are expected to

[7] See https://en.wikipedia.org/wiki/Brett_Kavanaugh#Sexual_assault_allegations for a concise description of the events

co-vary with voting intention), and extracted relevant tweets which fit into these topics, the next step is to train and tune machine learning classifiers.

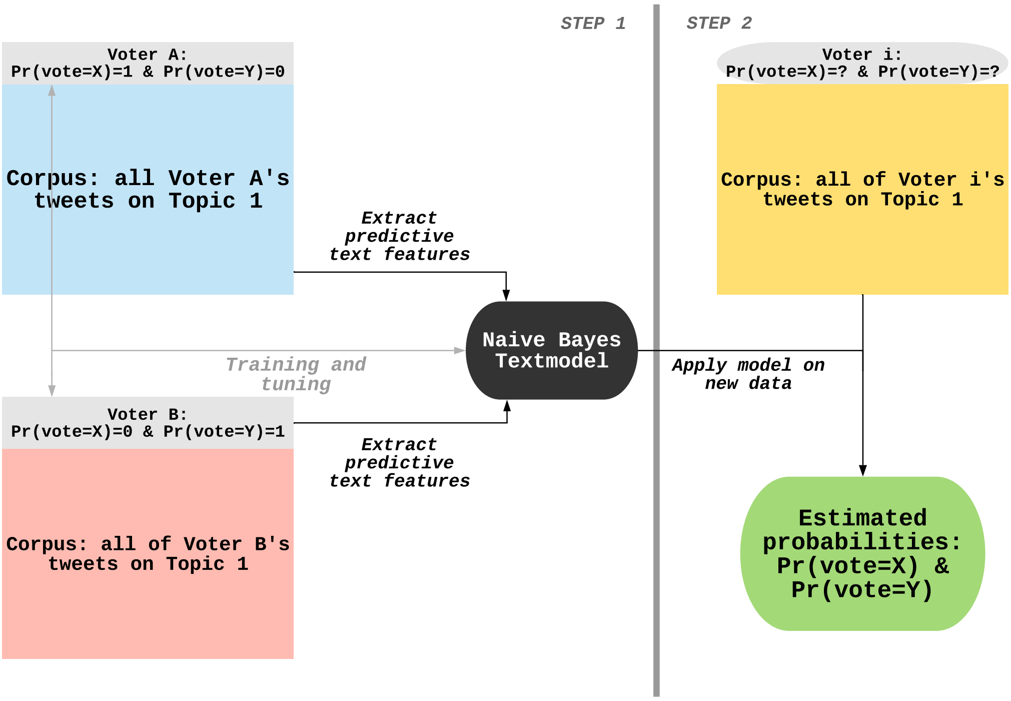

STEP 1
STEP 2
Voter A: Pr(vote=X)=1 & Pr(vote=Y)=0
Corpus: all Voter A's tweets on Topic 1
Extract predictive text features
Training and tuning
Naive Bayes Textmodel
Voter B: Pr(vote=X)=0 & Pr(vote=Y)=1
Extract predictive text features
Corpus: all of Voter B's tweets on Topic 1
Voter i: Pr(vote=X)=? & Pr(vote=Y)=?
Corpus: all of Voter i's tweets on Topic 1
Apply model on new data
Estimated probabilities: Pr(vote=X) & Pr(vote=Y)

Figure 6.2: The vote choice estimation process

This process is schematically depicted in the 'Step 1' segment of Figure 2[8]. Using the quanteda package for R (Benoit, 2018), relevant topic-level tweets for users whose voting intention was human-labelled are extracted and concatenated into corpora. These corpora are pre-processed, meaning the removal of stopwords and conversion to lower-case. Then, they are converted into document-feature matrices, denoting the frequency of each *feature* (aka word/uni-gram) in a given document (here, the corpus of concatenated tweets relating to a topic). The classifier is then trained by predicting the probability of the outcome variable (voting for X, or !X, or Y or !Y) on the training labels as a function of the observed frequency of features - words - in the document-feature matrix, the training data. These classifications are 'learned' by the algorithm, and applied on new, unlabelled data in order to predict the probability of the same, in this case unobserved outcome variable. However, before applying the trained classifier on new data, it is important to tune the classifier. This involves trimming document-feature matrices to adjust the number of documents and features given the predictive accuracy of the classifier when applied to known, previously labelled data. For this purpose, I extract a 20% holdout set from all corpora of labelled users' filtered topic tweets, and then split the remaining 80% into 10 equally sized samples, each of which are again held out of the training data for one iteration and used to evaluate the trained classifier's predictive accuracy. This process is repeated with multiple possible imputations of minimum and maximum features and documents in a document-feature matrix, while recording the model's accuracy parameters - precision, recall and accuracy. For each model, the tuning configuration with the best consensus performance between precision and recall

[8] For this paper, I exclusively employ the Naive Bayes machine learning algorithm, but other algorithms such as Support Vector Machines or Random Forests could also be applied to this problem

is chosen and applied to the 20% original holdout set. If the predictive accuracy is within +/- 5% of the previously observed levels, it is considered ‘ready’ to be applied on new data.

Applying models: Estimating individual-level vote choice

Having trained the models, the individual topic-specific classifiers can be applied on new, unlabelled data. For this purpose, the process of converting topic-level tweet corpora to trimmed document-feature matrices is repeated for ***all*** sampled users, not just those who tweeted their voting intention. Then, user-level probabilities of all desired outcome variables for which models were trained, in this case P(vote=Democrat) and P(vote=Republican) are computed. The validity of these computed probabilities is not only dependent on the availability of sufficient training data, but also on the availability of sufficient data to apply the classification algorithm to. If, for instance, a given user has only one relevant tweet, and the tweet features only a link and a two-word description of the link, it will not work.

I apply different classifiers for each of the topics outlined above - *Trump, Kavanaugh, Midterms, Republicans, Democrats*. Not every user produces relevant tweets for every topic, meaning that in several cases, the modelling process fails due to unavailable data. However, given that multiple topics were defined and multiple models trained, the number of computed user-classifications is maximised.

It is useful to investigate the substantive nature of what the different topic-level classifiers are *actually* measuring - conceptually, they measure user-level likelihood of vote choice for party X, but the noisy nature of Twitter data means that classifications from models applied on different data will likely lead to divergent classifications. This is best achieved by comparing user-level classifications across models. I do this by calculating inter-estimate Pearson correlations, which allow for a quantitative indication of how different models produce different classifications. The understanding of why this may be the case is a matter of interpretation.

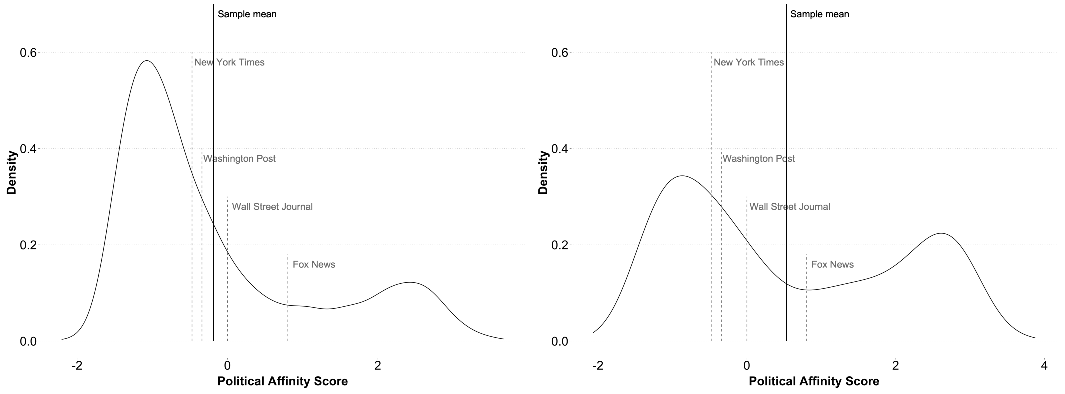
Furthermore, classifications are probabilistic, with each value denoting the probability of user i voting for party X, contingent on the available data. I define the cutoff value as *0.5*.

6.4 Data

I empirically test this framework for individual-level vote choice estimation in the context of the 2018 US midterms. US midterms occur every four years, two years after/before presidential elections. In them, the entire House of Representatives is elected, as well as a third of the Senate, and several state and lower-level executives and legislatures. When considering national offices, the midterms are a useful application for my proposed method, as they offer a scenario where several elections happen simultaneously, which are mostly (A) governed by marginally different but pre-defined electoral rules, while being contested by the same parties across states (B), and are understood as being determined by different factors than presidential elections (C). Furthermore, such factors are likely to differ to varying degrees across states, where different regional issues and underlying structural factors mesh with national issues in shaping eventual election results. For this paper, the goal is not to *forecast* the midterms simply for the sake of producing a refined model, but rather as a means of tracing public opinion on the midterms through Twitter. So, I am testing the applicability of my proposed method for estimating individual political preferences, and further to test the representativeness of resulting aggregations of these estimates in different sampling scenarios. Furthermore, the use of geo-located Twitter data allows for a geographically dis-aggregated analysis. Hence, I choose the empirical focus of this paper to be on the US nationally as well as its four most populous states: *California, New York, Florida and Texas*.

I sampled two distinctive sets of US twitter users. First, a 'civic panel' of Twitter users who tweeted that they voted in the 2016 US presidential election cycle. If previous voting is understood as an indicator of *civic duty*, these individuals are likely to vote in subsequent elections (Campbell et al., 1980, p. 105). This effect has previously been observed empirically (Matsusaka and Palda, 1999) and is thus operationalised for the purpose of sampling users who are (A) likely to tweet politically salient content necessary to estimate their voting intention and (B) likely to turn out. Users were sampled by filtering all pertinent, accessible tweet collections at NYU's SMaPP lab [9] concerning the 2016 US presidential election for keywords indicative of voting: '#Ivoted', 'I voted for', '#myvote2016', '#myvote', '#Ivotedbecause'. This left **63,400 unique tweets**. I removed tweets with verbatim duplicate text (a typical indicator of bots) and retweets, as they cannot be understood as a clear indication of an individual's vote, but rather of their endorsement of another person's vote. I then extracted all unique user ids, leaving a total of **13,928 unique users**. I then geo-located all users using my pipeline for geo-locating Twitter users (Loynes et al., forthcoming), leaving a total of **5,164 unique users** which could unambiguously be placed in the United States at the state-level.

Second, I use a sample of 30,000 randomly selected Twitter users[10], who were geo-located using the same method. Users not in the US were discarded, resulting in a sample of **10,000 unique US users**. I refer to this sample as the 'control panel', as it is useful for the purpose of testing my methodological approach on distinctly different samples in order to investigate the efficacy of the method given different data and to investigate factors contributing to inter-sample variation.



(a) Civic panel

(b) Control panel

Figure 6.3: *Tweetscores* ideal-point estimate distributions

I applied ideal-point estimates for political affinity to all users in both panels using the 'tweetscores'[11] algorithm (Barberá, 2015). As shown in Figure 3, the two samples differ greatly

[9] Time frame of collections: January 2016 to November 2016, thereby including a large number of state-level primaries

[10] A random number generator was used to create Twitter user-ids

[11] In essence, tweetscores leverages a given user's following-network to estimate their latent political ideology

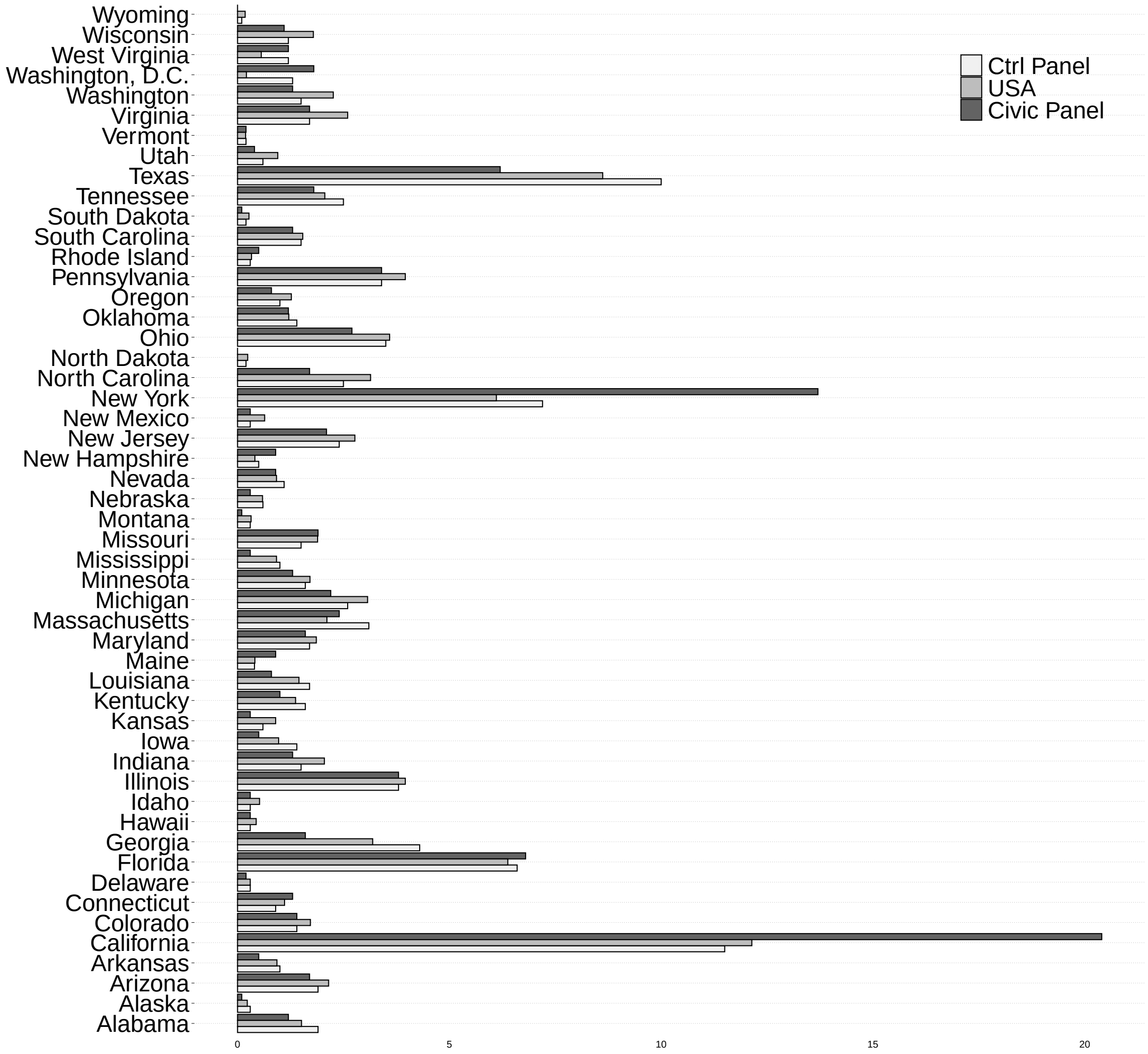
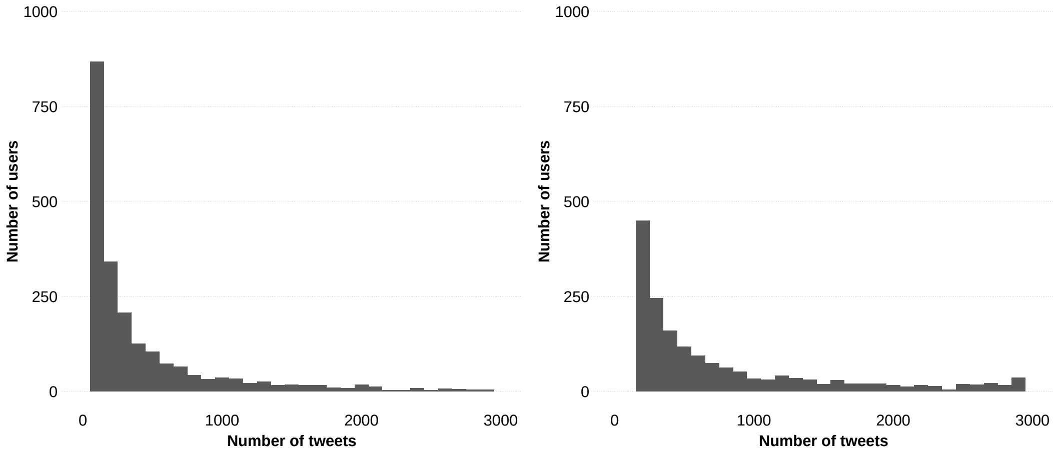


Figure 6.4: Location distribution - samples & real world

in regard to their partisan composition. The mean user in the *civic panel* is significantly further to the left (-0.186) than in the *control panel* (0.531), and further, the control panel features a considerably more balanced partisan composition with two closer modal peaks on both the left and the right, and a larger proportion of what might be referred to as 'moderates', i.e. users located between Fox News (0.803) and the New York Times (-0.472). Conversely, the civic panel is heavily biased to left-wing users. This may be an indicator that left-wing Twitter users are more likely to share their act of voting on Twitter than right-wing users, but may also show that left-wing users are more common in the 'political Twitter' space than right-wing users. Furthermore, it is important to note that not all 5,164 users (civic panel) / 10,000 users (control panel) were classifiable using *tweetscores*, as the algorithm relies on users following at least 3 ideology-relevant 'elite' accounts. For the civic panel, **3115** users were classified, while the control panel only contained **2614** classifiable users. This further supports the idea that 'political Twitter' is dominated by more left-wing users.

under the assumption of homophily. Certain 'elite' accounts, such as politicians, pundits/journalists or news organisations can be ideologically classified. A target user's tweetscore is then determined by which of this pre-defined elite accounts they follow.

Figure 4 depicts the proportion of users from either panel as well as the real-world US Census (2010) proportion in all 50 states and Washington, D.C. It is immediately obvious that users from California and New York - both considered liberal states - are significantly over-represented in the civic panel when compared to the US population. The same is true, albeit to a less pronounced extent, for Texas and the *control panel*. This further suggests that the civic panel is biased in a leftward direction, while the control panel is more balanced and, on the whole, closer to representing the offline world, at least when it comes to baseline partisanship or location. Generally, the plot indicates that for several less populous states, both panels are under-sampled compared to the actual US population. Overall, however, it is important to note that, excluding California and New York, both samples include users from all states, and are mostly not radically divergent from proportionate geographic representation of the real world.



(a) Civic panel

(b) Control panel

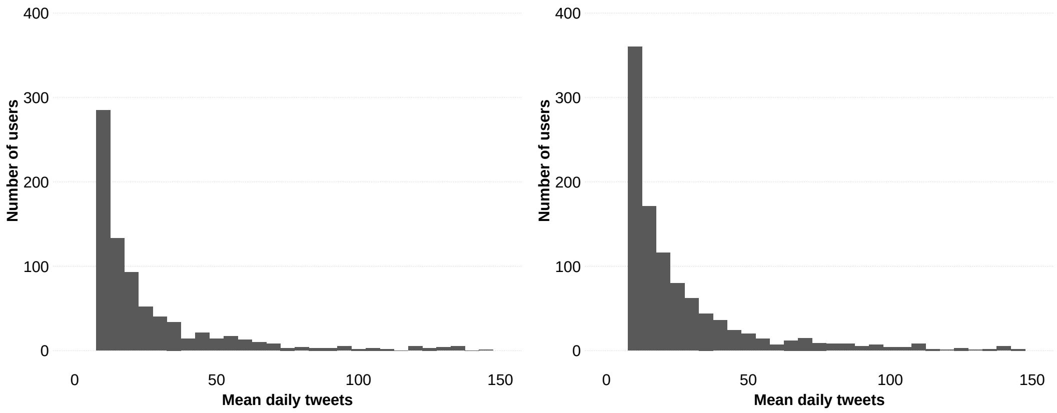
Figure 6.5: Total number of tweets (06 Sep 2018 - 05 Nov 2018)

In order to build models to estimate individuals' vote choices, I collected *all* tweets - regardless of expected political salience - published by members of either panel in the time from September 6th, 2018 to November 5th, 2018 using the rtweet package for R (Kearney, 2018).

Given how the civic and control panels differ in regard to their users' political affinity and their users' locations, it is useful to also investigate their tweeting behaviour. Figure 5 depicts histograms for both panels' users' total numbers of tweets. The control panel features fewer users who do *not* or only rarely tweet than the civic panel, and more users who tweet a lot, with noticeable bins at the right end of the distribution. The mean total number of tweets for the civic panel, 265, is considerably higher than for the control panel, 165.

Figure 6 shows histograms of the mean number of daily tweets for both panels. The graphs indicate that control panel users have a higher number of daily tweets than civic panel users, but, again, there is a large degree of variance and the modal category is people who tweet between 0 and 1 times a day, for both panels. The similarity in both samples' distributions suggests that both samples feature mostly users whose tweeting/sharing patterns are not hugely different.

In summary, a brief investigation of relevant descriptive statistics of both the civic and control panels indicates that the samples are distinctly different in their socio-demographic



(a) Civic panel

(b) Control panel

Figure 6.6: Mean daily tweets

composition, strongly reflected in the distribution of tweetscores and users' home locations. While this is likely explainable in part by the sampling strategy employed, it is impossible to conclusively determine this without comparing these samples to further, potentially larger and differently generated samples of Twitter users. For the purpose of this paper however, the difference between both samples allows for a useful test and comparison of the breadth of the applicability of my proposed method.

6.5 Results

Before discussing findings, it is useful to look at model specifications and models' applicability for the task at hand, as well as the performance of the models.

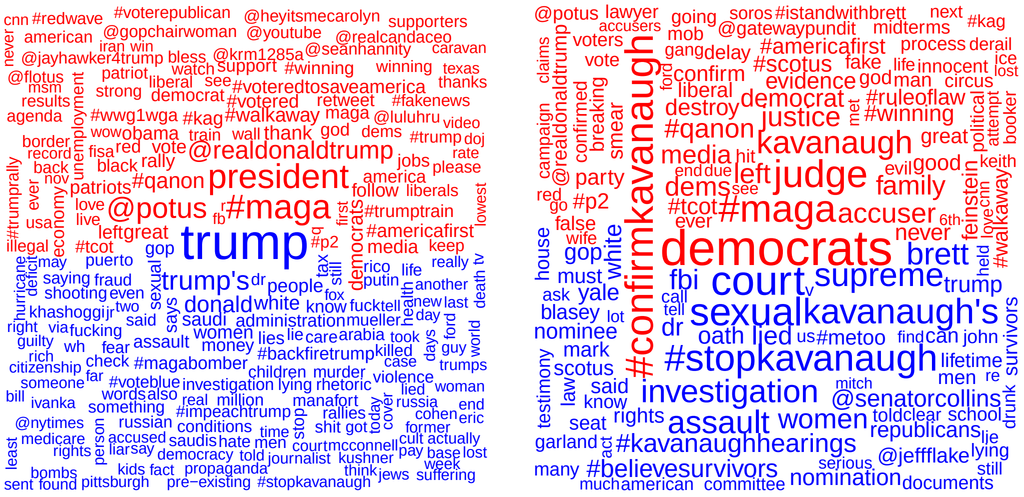
6.5.1 Model specifications and performance

As outlined above, it is first necessary to extract relevant ground truth data and hand-annotate these data according to the substantive signals contained in pertinent tweet text. Using the keyword-based approach outlined earlier, I annotated n=1586 tweets (civic panel) and n=1730 tweets (control panel), whereby the annotations reflect tweet-level voting intention. These annotations were then collapsed to the user-level, as several users had multiple tweets in the ground truth sets, resulting in ground truth sets of ***n* = 372 users** (civic panel) and ***n* = 391** users (control panel).

	Democrats		Republicans		*other*	
	n	%	*n*	%	*n*	%
Civic Panel	231	62%	141	38%	-	-
Control Panel	97	25%	227	58%	67	17%

Table 6.1: Ground truth voting intention: distribution by sample

Table 1 shows the distribution of ground truth labels for both samples by hand-labelled user-level voting intention. This provides further evidence for the politically different composition of the two samples, whereby each subset of ground truth 'voters' leans towards the opposite party with a similar weight.



(a) Trump model

(b) Kavanaugh model

Figure 6.7: Wordclouds: Most frequent terms by users' ground-truth party preference

As this research is based on differences in how people discuss politics as a function of their observed party preference, Figure 7 shows wordcloud plots comparing the most frequently occurring terms in the hand-annotated ground truth training data for the civic panel[12]. The larger the text, the more frequently words are contained in the tweet text posted by gold standard set users in a given topic. The red and blue highlights signify words used by Republican and Democrat voters respectively. Both plots show that Democrat and Republican voters talk about model-relevant topics, in this case *Trump* and *Kavanaugh*, exceedinlgy differently. Moreover, the figures contain terms which relate closely to sub-issues which were relevant to both current events and the election campaign. This is particularly clear for the Kavanaugh model, where the ideological difference between opposite partisans manifests itself in the most frequently used hashtags: *#stopkavanuagh* and *#confirmkavanaugh* respectively. However, wordclouds, while functioning as a useful tool to aid interpretation, offer no robust measure of difference between underlying text corpora associated with either party preference. For this purpose, I further computed several distance metrics for all models for both samples, displayed in Table 2.

Table 2 contains measures for cosine similarity, Pearson correlation and Euclidian distance between the sub-groups (Democrat/Republican voting intention) in the ground truth data for both samples. For cosine similarity, a value closer to 1 indicates higher similarity of text corpora. Hence, users who tweeted their voting intention for either Democrats or Republicans use exceedingly distinct terms when talking about any of the five related topics. For "Distance", a higher value indicates lower similarity between terms used by the users in the labelled set when tweeting about relevant topics. In this case, this suggests that the midterms model is likely least appropriate for this task, as, compared to the other models, the terms used for either group are closest to one another on average. The computed distance metrics also suggest that the Trump model is highly suitable for the task at hand: in both

[12] See the appendix for wordcloud plots for the other models

samples, the terms used by voters of either party are the most far away from one another when compared to the other models in the same sample.

Model	Civic Panel			Control Panel		
	Cosine sim.	Correlation	Distance	Cosine sim.	Correlation	Distance
Trump	0.00	-0.05	589.52	0.00	-0.02	1691.05
Kavanaugh	0.00	-0.04	443.81	0.00	-0.04	523.66
Midterms	0.00	-0.11	153.80	0.00	-0.11	178.99
Democrats	0.00	-0.08	333.22	0.00	-0.07	475.70
Republicans	0.00	-0.05	374.98	0.00	-0.02	1316.82

Table 6.2: Text similarity between Democrats and Republicans, weighted by inverse document frequencies

Having established the *a priori* suitability of all models, it is further useful to investigate each models' classification performance. I computed three measures of model performance: accuracy, precision and recall. Accuracy signifies the proportion of correctly classified cases out of all classifications, or the sum of true positives and true negatives divided by all classifications. Precision refers to the number of correctly classified positive cases (true positives) divided by all positively classified cases (true positives and false positives), while recall is the quotient of all correcty classified positive cases (true positives) over all positive cases (true positives and false negatives). Machine learning models can be tuned so as to achieve optimal values in any of the above measures by adjusting model parameters, such as - in this case - the minimum/maximum term and document frequencies of the training corpora. However, it is at the researcher's behest which dimension of classification they prioritise, which, by definition involves trading off an increase in precision for a decrease in recall, and vice versa. In this case, I tuned models in order to maximise both precision and recall.

Model	Civic Panel			Control Panel		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Trump	0.950	0.957	0.960	0.882	0.700	0.939
Kavanaugh	0.936	0.995	0.901	0.894	0.772	0.869
Midterms	0.826	0.980	0.731	0.879	0.768	0.765
Democrats	0.893	0.965	0.859	0.895	0.759	0.841
Republicans	0.916	0.956	0.901	0.885	0.745	0.820

Table 6.3: Model classification performance measured on 20% holdout sets

Model performance was computed through 10-fold cross-validation paired with hold-out validation. To begin, a random 20% sample was drawn from the ground truth samples and set aside. Then, the remaining data were split into 10 equally sized sub-samples, each of which was used as a test set once with 441 different model hyper-parameter configurations. Finally, having determined the 3 best-performing parameter configurations, I tested model performance on the initial 20% holdout set. Table 3 depicts the performance metrics of specified models, for both samples, as measured on the 20% holdout set.

The performance figures for all models in both samples considerably exceed the baseline value of 0.5. All models provide strong performance in both samples. However, this research design provides the unique opportunity of further broadening the scope of validation to truly distinct, previously unseen data. For this purpose, I validated each trained model on the ground truth data from the respective other sample, i.e. control panel classifier on civic panel data, and vice versa. Given the different composition of the samples, and especially their ground truth sets (see Table 1), this is a useful test to assure the accuracy of the model specification. Table 4 shows the performance metrics for this out-of-sample validation, as well as the number of cases which were not classified. All models' performance is again very high. However, precision levels for civic panel models on control panel data are lower than in the previous validation stage. Overall, control panel models on civic panel data perform better than the inverse category of models. This is likely due to the larger degree of variance in the ideological composition of the control panel vs the civic panel (see Figure 3).

	Model	Accuracy	Precision	Recall	*n* unclassfied
Civ -> Ctrl	Trump	0.89	0.70	0.97	19
	Kavanaugh	0.90	0.74	0.94	35
	Midterms	0.91	0.74	0.95	67
	Democrats	0.91	0.75	0.95	42
	Republicans	0.87	0.67	0.94	27
Ctrl -> Civ	Trump	0.97	0.96	0.99	7
	Kavanaugh	0.97	0.99	0.97	21
	Midterms	0.95	0.97	0.95	69
	Democrats	0.94	0.97	0.93	28
	Republicans	0.96	0.98	0.96	21

Table 6.4: Cross-sample classification performance

6.5.2 Classifications

Table 5 shows the results of all the number of cases classified as one of the two parties (V=D, V=R and V=D*) when selecting a cutoff value of 0.5 for each classified predicted probability[13]. Furthermore, Table 5 shows the percentage of users classified as likely Democrat or Republican voters (%D, %R, %D*) out of all classified users for that particular model, *not* the entirety of the given sample's *n*. Finally, the "*n* NA" column indicates the number of users who were not classifiable by a given model.

First, it is interesting to note the stark divergence in party-level classification share between the two samples. As may be expected given the measured ideological composition of the two samples (see '5. Data'), the Democratic classified vote share for the civic panel ranges between 83% and 68%, depending on the specific model. The highest-performing models (Trump, Kavanaugh and Republicans) produce classified vote shares between 75% and 83% for the Democrats. The control panel classifications are more balanced between the two parties, ranging between 65% - 43% Democratic vote share depending on the

[13] The classifier returns a predicted probability of the classified case being the positive value. While it can be useful to fine-tune a classification cut-off point different from 0.5 in the validation stage of model training, this should be done on a case-by-case basis, and given the high performance values for most models observed for these models, I forgo this step.

model, whereas the best-performing models still classify a majority of the sample as likely Democratic voters.

	Model	V=D	%D	V=R	%R	*n* NA	V=D*	%D*
Civ	Trump	1964	83.11	408	17.27	2813	2159	79.14
	Kavanaugh	1110	75.41	384	26.09	3704	1274	69.88
	Midterms	529	79.43	137	20.57	4510	635	65.53
	Democrats	713	68.56	336	32.31	4136	858	61.99
	Republicans	1061	74.56	383	26.91	3753	1213	68.38
	Mean	2241	81.05	539	19.49	2411	2379	75.91
Ctrl	Trump	932	65.27	412	28.85	8350	1126	62.56
	Kavanaugh	551	56.17	440	44.85	8797	693	51.83
	Midterms	263	51.98	236	46.64	9272	384	46.27
	Democrats	291	43.43	383	57.16	9108	427	41.90
	Republicans	485	50.84	463	48.53	8824	684	51.90
	Mean	1185	63.20	640	34.13	7903	1439	63.73

** cross-sample classification, i.e. Civ classifier on Ctrl data and vice versa*

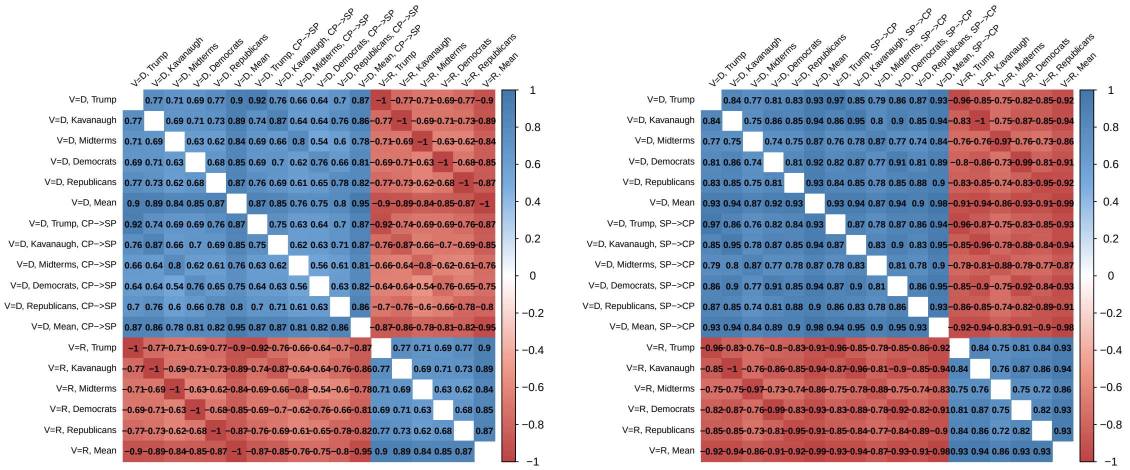
Table 6.5: Vote-choice classifications

Second, the two samples differ greatly in the proportion of classifiable users. For the civic panel, the Trump model features the highest coverage (2,372 users out of 5,176), while the same is true for the Trump model in the control panel of randomly selected users (1,344 users out of 10,000). I argue that the fundamentally different composition of these samples regarding their members' relative 'political-ness' can account for much of this difference: people who take voting so seriously that they publicise their doing so are more likely to tweet about politically salient topics than the average, randomly selected US Twitter user. The findings shown in Table 5, as well as the tweetscores distribution further indicate that such users not only tweet more about politics, but are also considerably further to the left than the average US Twitter user.

Besides the pre-specified models, Table 5 also shows classification metrics for user-level mean predicted vote choice probability across all available model classifications. In other words, if a user is classified by 3 distinct models, the 'Mean' value for this user is the mean of these predicted probabilities. Hence, the 'Mean' of all available models maximises the number of classified users to be analysed en-bloc, as it combines all available classifications. However, the mean does not take into account the relative performance, both conceptually and empirically, of the models, meaning that bias may be introduced through artifacts of lower-performance or inadequately specified models.

In order to evaluate how the different models compare in their classification of vote choice probabilities, and further, to evaluate the usefulness of the 'Mean' model, I computed Pearson correlations between all models' classifications for both samples, as displayed in Figure 8. Inter-model classification correlations are high for all models and samples. This is also the case for correlations between classifications obtained from cross-sample and intra-sample classifications (i.e. employing the control panel classifier on civic panel data and vice versa).

As may be expected given the model performance metrics discussed in the above section,



(a) Civic Panel

(b) Control Panel

Figure 6.8: Inter-model Pearson correlations for classified vote-share probabilities

the "Trump", "Kavanuagh" and "Republicans" models' classifications share the highest correlations, regardless of which classifier was used, and for both samples. The other two models have lower correlation coefficients ranging below the .75 - range, but nonetheless in ranges that would be considered high to moderately high in traditional social science research. Furthermore, the mean classification value is very highly correlated with all models, for both samples, suggesting that this method of aggregating all available classifications is a useful way of maximising coverage and generalising the predictive power of all models to a single metric for all available users.

Finally, there is the question of which user-level characteristics predict vote choice. For this purpose, I fitted several logistic regression models, with user-level ideal-point ideology estimates (Barberá, 2015) and demographic estimates (race and sex) as predictors. Users' sex was estimated using the *gendeR* package for R (Mullen, 2018); race was estimated using the *wru* package for R (Imai and Khanna, 2016). Both packages match names to US census records and compute probabilities of a name being associated with a sex / race category given the frequency with which it appears in the census. This method is probabilistic and noisy, as several users do *not* provide clear first name/last name pairs in their Twitter profile information. Hence, these data are presented for illustrative purposes only and results should be treated with a degree of caution. Table 6 shows the regression table for mean model-derived classifications - Democrat vote, Republican vote, and Democrat vote in the cross-sample classification for the Civic Panel[14].

Political ideology is a strong predictor of voting intention across models and samples. The lower the tweetscore, the further left a user is placed on the ideological spectrum. The demographic variables have little to no significant effect on vote choice for the data at hand, apart from a moderate statistically significant measured association between being black and a higher propensity to be classified with a likely vote for the Democrats. However, in the civic panel this is only the case for the cross-panel classification model. This is likely

[14] See Appendix B for the equivalent table for the Control Panel

	Dependent variable:		
	V=D	V=R	V=D, Civ->Ctrl
	(1)	(2)	(3)
Tweetscores ideology	-1.045^{***}	1.024^{***}	-1.094^{***}
	(0.067)	(0.066)	(0.059)
Sex: male	−0.248	0.258	−0.009
	(0.176)	(0.173)	(0.156)
Ethnicity: asian	−0.118	0.040	0.119
	(0.490)	(0.487)	(0.464)
Ethnicity: black	1.005	−1.066	1.553^{*}
	(0.813)	(0.809)	(0.806)
Ethnicity: latin/hispanic	−0.502	0.422	−0.405
	(0.350)	(0.346)	(0.319)
Ethnicity: other	11.914	−11.012	−0.939
	(620.249)	(376.098)	(1.364)
Ethnicity: white	0.303	−0.309	0.175
	(0.219)	(0.215)	(0.193)
Constant	1.615^{***}	-1.546^{***}	1.295^{***}
	(0.208)	(0.204)	(0.181)
Observations	1,292	1,292	1,502
Log Likelihood	−441.520	−454.588	−553.639
Akaike Inf. Crit.	899.040	925.177	1,123.277
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 6.6: Logistic regression models predicting classified user-level vote choice (Civic Panel)

due to both the uncertainty regarding these demographic estimation algorithms and their specific shortcomings when applied to non-standard Twitter profile data, but likely also stems from the unrepresentative nature of both samples when compared to the population of (offline) US adults. However, it is clear that the strong effect of computed ideal-point ideology estimates and the vote choice classifications computed using the method described in this research means that there is a strong internal consistency between the method's rationale and its empirically observed relationship between previously verified, computed user-level characteristics, which are also inherently theoretically linked with the act of voting - while other factors, such as the state of the economy influence individuals' voting decision, this is often seen as being superseded by individuals' baseline ideological affiliation, which, in a two-party system like the USA's, can often be equated to party identification.

When using individual-level vote choice classifications as a building block for estimating aggregate public opinion, these logistic regression models serve a further purpose: by using the logistic regression models to predict their outcome variables for previously unclassified cases using the available dependent variables, the number of cases classified for likely vote choice in any sample can be maximised. For the purpose of evaluating model performance by aggregating vote choice classifications, I also compute these metrics with the logit model method.

6.5.3 Aggregating individual-level vote-choice classifications: Post-casting the 2018 midterms

Finally, we want to see how individual-level vote choice classifications can be aggregated to the national and sub-national levels. It is important to note that the goal of this research is not to produce the best possible 'post-cast' of the 2018 US midterms, but rather to demonstrate how user-level public opinion can be estimated, and how identically specified models perform on different samples of users. However, it is still useful to compare aggregations of vote-choice classifications with actual election results in order to see how closely the different samples come to approximating the offline reality, i.e. the elections' outcomes. The insights gleaned from this exercise can then be used when applying this method with new data in the future, especially if the goal *is* forecasting elections.

I focus on congressional popular vote percentages for both parties in the national context, as well as the four most populous states - California, Texas, New York and Florida as benchmarks to compare aggregated vote choice classification tallies to. I choose not to analyse the remaining 46 states as the number of sampled users in either sample would not suffice to draw any broader conclusions for less populous states. However, assuming this method were applied to a large sample of Twitter users from e.g. Alabama, there is no obvious reason why it should not be able to produce an accurate representation of the eventual election result, assuming correct sampling.

I obtained Congressional-district-level vote tallies in the 2018 US midterm elections from the "2018 House Popular Vote tracker" spreadsheet compiled and shared by the Cook Political Report (Wasserman and Flinn, 2019). In order to obtain state-level House popular vote figures, I summed the total number of votes in each of the relevant states for each party, and calculated percentages.

Besides presenting the raw vote choice classification tallies for the national level and the 4 states for several models, I also present classification percentages weighted by relative population proportion of the individual states. In order to do this, I first calculated a weighting factor, i.e. the ratio of over- or under-representation of the number of users in a given sample as a proportion of the entire sample compared to the given states' real-world proportion of the country's population as a whole. Then, I drew bootstrapped sub-samples of users for a given state, so that the weighted number of users from that state is proportional to the state's actual population proportion of the US as a whole. This was repeated 2000 times for each state, and the mean number of vote classifications for either party across all 2000 iterations were recorded.

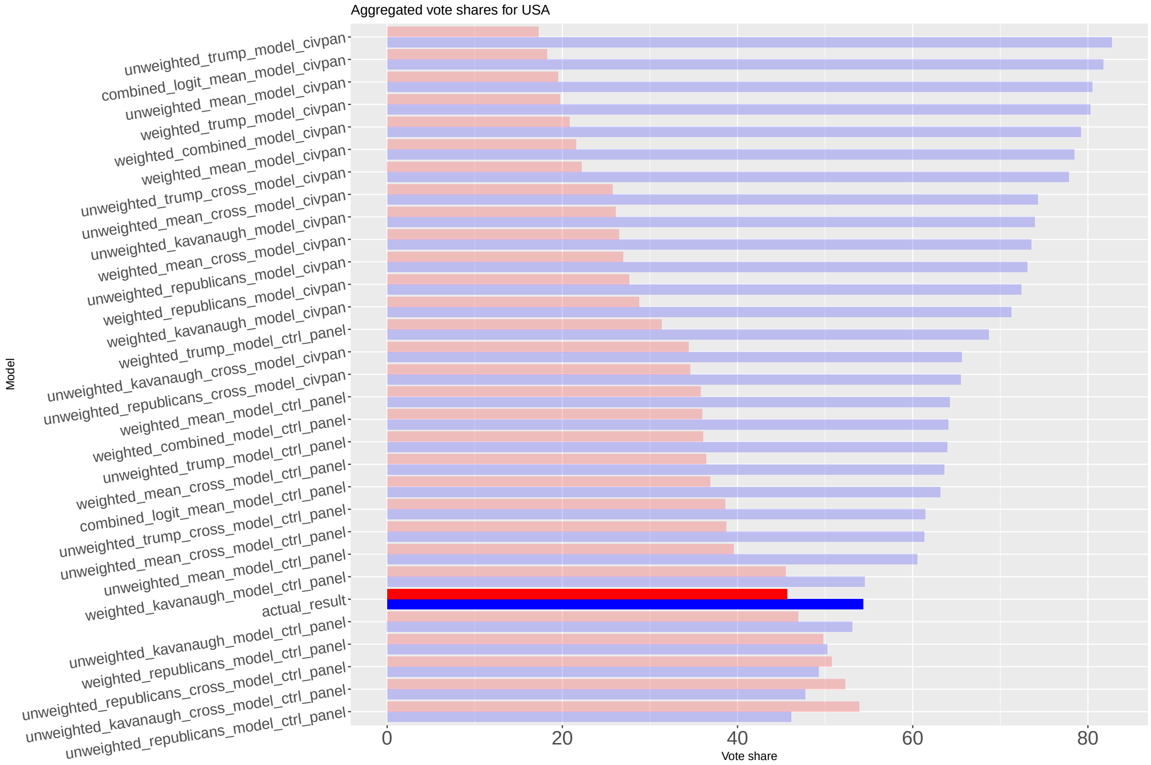


Figure 6.9: Aggregate vote shares: model predictions and actual result, national congressional popular vote

Figure 9 shows the predicted vote shares for the US national congressional popular vote, for models in both weighted and unweighted configurations, for both the on-sample classifications as well as cross-sample classifications, and user-level means across all available models. 27 out of 30 models predict the correct winning party of the popular vote, the Democrats. However, the vast majority of models heavily overestimate the Democrats' vote share while underestimating the Republicans'. The respective bars for each model's predicted vote share are displayed in descending order by the size of prediction error compared to the actual election result ("actual_result").

The weighted Kavanaugh model applied on the control panel provides the most accurate prediction of the eventual result, which, at 54.5% for the Democratic Party and 45.5% for the Republican Party is approximately 0.15% away from the actual result, 45.64% (Republicans) and 54.36% (Democrats). This low error exceeds pre-election polling in terms of predictive accuracy, and considerably exceeds the predictive accuracy of the majority of previously published Twitter-based election forecasts. The unweighted Kavanaugh model, as well as the weighted Republicans model also come very close to the actual election result. Furthermore, it is apparent that across models, aggregated vote shares obtained from the control panel significantly outperform those obtained from the civic panel, which is likely explainable by the ideological composition of the sample. Only 7 models are within a 5% error distance of the actual result, while the remaining 23 models have an error range between 6% and 29%, which implies that these models are not useful for the purpose of using aggregated classifications to gain insights into offline public opinion.

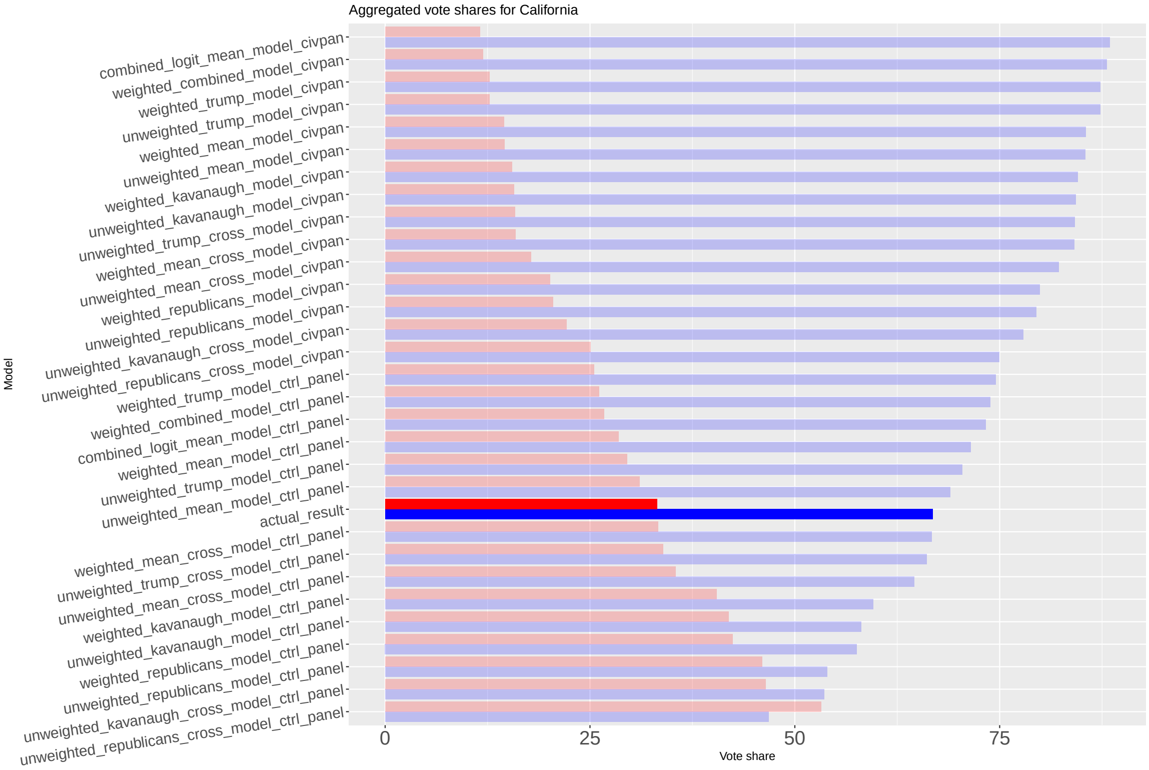


Figure 6.10: Aggregate vote shares: model predictions and actual result, California congressional popular vote

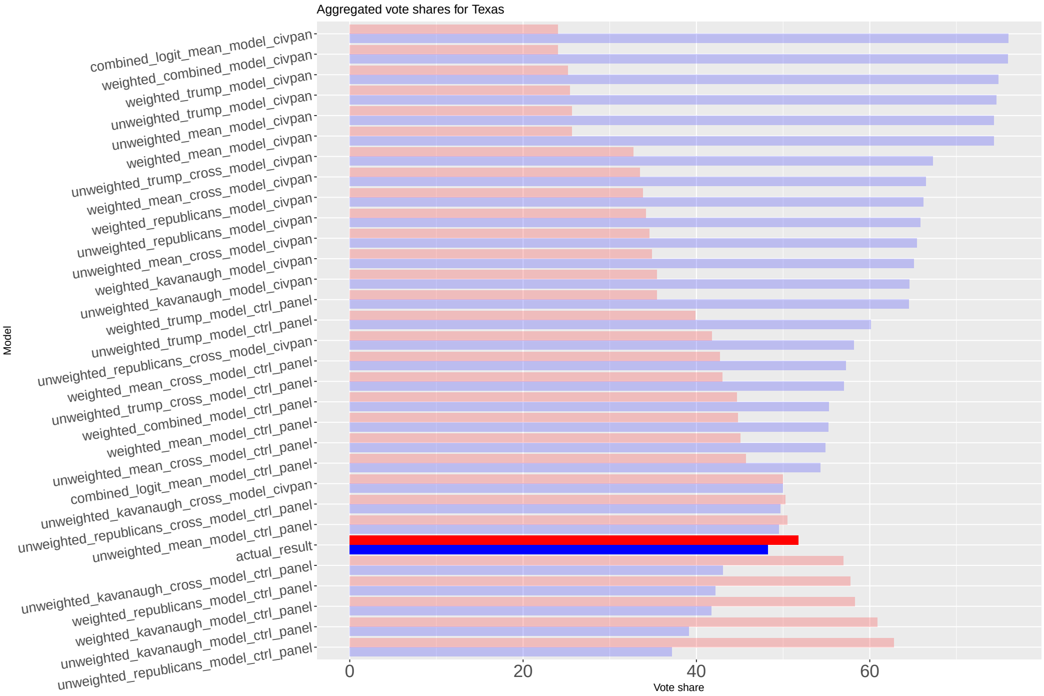


Figure 6.11: Aggregate vote shares: model predictions and actual result, Texas congressional popular vote

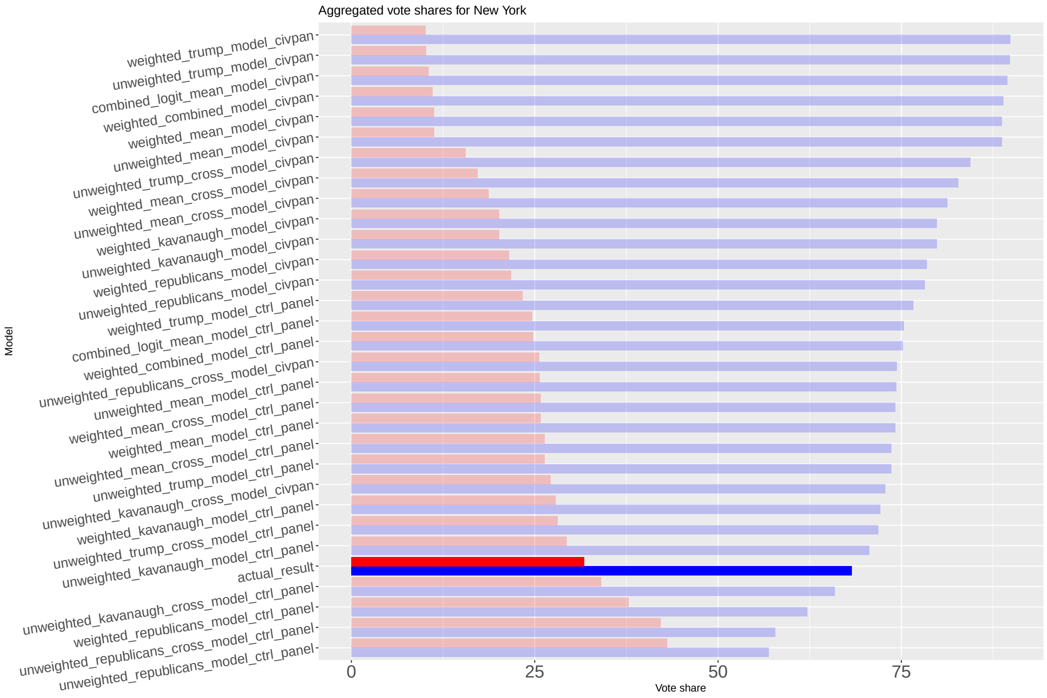


Figure 6.12: Aggregate vote shares: model predictions and actual result, New York congressional popular vote

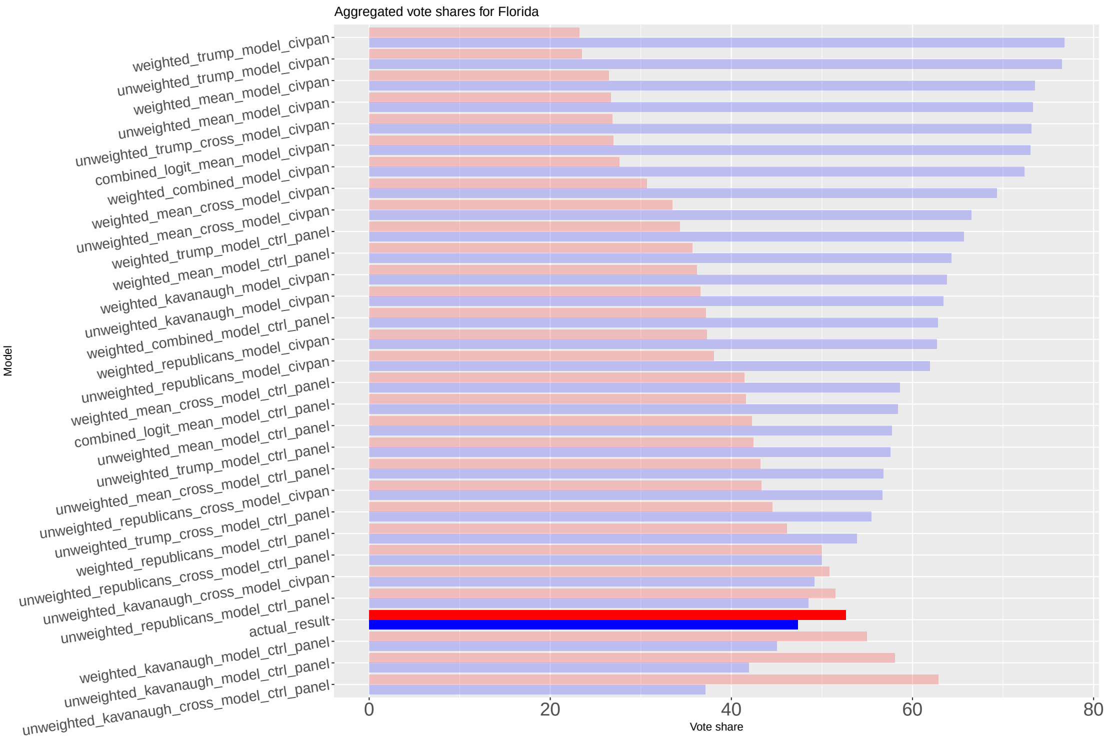


Figure 6.13: Aggregate vote shares: model predictions and actual result, Florida congressional popular vote

Figures 10 through 13 show the equivalent vote share distributions for models and House popular vote tallies for the states of California, Texas, New York and Florida. In the case of California, 29 out of 30 models correctly predict the party with the majority in the popular vote; for Texas, 7 out of 30 models predict the majority party; for New York, all models predict the Democrats as popular vote winners; while for Florida, 5 out of 30 models get the popular vote winner right. For all models, the Kavanaugh and Republicans models, in both weighted and unweighted specifications, again perform best at classifying the control panel users' likely vote choice in line with the eventual result. Furthermore, it is again clearly apparent that aggregated classifications from the control panel reflect state-level popular vote tallies better than those from the civic panel. Finally, it is clear that aggregated classifications generated by the majority of models - regardless of which of the two samples they were generated on - are biased towards the Democrats. It seems plausible that this is due to the fact that liberal-leaning users are simply more common on Twitter than conservative-leaning users, and thus - regardless of post-stratification with population weights - the models classify *most* users as Democrats, because most users in either sample *are* Democrats.

In order to comparatively evaluate each model's average performance when aggregating its individual-level vote choice estimates to the 5 specific popular vote tallies, I computed the mean error (in percentage points) between aggregated vote share totals produced by the classification models and the actual state- and national-level popular vote tallies. I calculated these distances by extracting the squared error for each model (so as to account for both over- and under-estimation) on each aggregation versus the actual result, and then taking the square root of the mean of all classification error for each model. Table 7 shows these computed distance scores in descending order. As indicated above, all control

Model	Mean distance (%)
weighted_kavanaugh_model_ctrl_panel	4.82
unweighted_mean_model_ctrl_panel	6.13
unweighted_kavanaugh_model_ctrl_panel	6.27
unweighted_trump_cross_model_ctrl_panel	6.43
unweighted_mean_cross_model_ctrl_panel	6.57
weighted_republicans_model_ctrl_panel	6.60
unweighted_kavanaugh_cross_model_civ_panel	7.43
combined_logit_mean_model_ctrl_panel	8.08
weighted_mean_cross_model_ctrl_panel	8.11
unweighted_kavanaugh_cross_model_ctrl_panel	8.46
unweighted_trump_model_ctrl_panel	8.72
unweighted_republicans_cross_model_civ_panel	9.08
weighted_combined_model_ctrl_panel	9.80
unweighted_republicans_model_ctrl_panel	9.84
weighted_mean_model_ctrl_panel	9.90
unweighted_republicans_cross_model_ctrl_panel	10.43
weighted_trump_model_ctrl_panel	13.68
unweighted_republicans_model_civ_panel	15.04
weighted_republicans_model_civ_panel	15.21
weighted_kavanaugh_model_civ_panel	16.06
unweighted_kavanaugh_model_civ_panel	16.41
unweighted_mean_cross_model_civ_panel	17.11
weighted_mean_cross_model_civ_panel	18.41
unweighted_trump_cross_model_civ_panel	20.68
weighted_mean_model_civ_panel	23.30
unweighted_mean_model_civ_panel	23.68
weighted_combined_model_civ_panel	24.02
combined_logit_mean_model_civ_panel	24.88
weighted_trump_model_civ_panel	25.01
unweighted_trump_model_civ_panel	25.43

Table 6.7: Mean model distance

panel models perform better than civic panel models, whereas civic panel-classifier cross-model classification on the control panel performed moderately well (error between 7.43 and 6.43 %) for the Trump, Kavanaugh and mean models. The weighted Kavanaugh model applied on the control panel results in the most accurate vote share predictions across all cases, but overall, for all of the best performing models, there is no indication that weighting necessarily improves model performance. Furthermore, the enhanced approach incorporating extrapolated vote share estimates by predicting previously unclassifiable users with logistic regression models proved not useful for the task of aggregating vote shares in the samples it was tested on, as a multitude of other models' aggregated vote share predictions proved more predictive of the eventual vote share result. This suggests that the noise introduced from computed demographic variables coupled with the noise introduced from uncertain logit-based classifications is not on par with tweet text-derived classifications.

In summary, the vast majority of aggregated vote shares obtained from classifying individual-level vote choice correctly predict the party with the majority share of the popular vote in the case of the 2018 US midterms, nationally, and for the 4 largest states. The weighted Kavanaugh model performs best across the board, while the civic panel is not suitable as a mirror of offline public opinion.

6.6 Discussion

This paper had several objectives: first, to introduce and clearly specify a novel methodological framework for estimating public opinion of Twitter users. Second, to test this method on distinctly different samples of Twitter users in the context of the 2018 US midterm elections. Third, to aggregate individual-level vote choice estimates to relevant geographic entities and thereby produce predictions for party level popular vote percentages, and to evaluate these aggregations in comparison to the real-world popular vote shares. Overall, this research achieved all the stated goals with encouraging findings, forming a solid foundation for future research.

First, the proposed vote choice classification method performs well. Across samples, several models perform with high to moderately high degrees of accuracy and provide robust classifications across users from different locations, with polarised political ideologies and different underlying socio-demographic attributes. The method performed well for both samples, whereas a significantly higher proportion of users were classifiable in the civic panel than the control panel. Most likely, this is due to the fact that the civic panel includes users who tweeted their voting intention 2016 US election cycle, whereas the control panel consists of randomly selected US Twitter users. This means that the civic panel produced more politically salient content, which is required for classifying users' likely vote choices.

The difference in numbers of classifiable users, as well as the distribution of classifications in both samples highlights one of the key takeaways of this research: *sampling decisions are crucial when measuring public opinion using tweets*. On average, the control panel of randomly selected US Twitter users is considerably more suitable as a lens into US public opinion than a sample of very politically engaged US twitter users. This indicates that tweeting *#imvoting* and related hashtags is an activity not evenly distributed across the ideological spectrum, and certainly not equally distributed across the population of US twitter users. Notwithstanding the encouraging findings obtained from the control panel, further research into a way of obtaining a sample which more accurately reflects target offline populations is necessary. This will involve new approaches of estimating latent individual-level characteristics, such as age, education and socio-economic status. Such variables, combined with other computed attributes will allow for a deeper investigation of the sources of bias in vote choice classification, and enable for the construction of specific large-n samples on which repeated, regular vote choice classification can be performed.

Furthermore, there is the question of how suited this method is to producing accurate election forecasts. While the goal of this paper was explicitly *not* forecasting the midterms, I nonetheless aggregated vote choice classifications obtained from a range of different model configurations. For the control panel of randomly selected twitter users, the Kavanaugh model predicted the popular vote party shares on par with pre-election polling across states. However, most other models fell considerably short, with mean error across cases ranging up to 25.34%. For the civic panel, this is easily explained by the sample composition. For combined logit and classification models, it is explainable by the compound noise. For the badly performing pre-specified models used on the control panel, I suggest it may be due to the specification of the models in relation to the theoretical model of political tweeting and vote choice. President Donald Trump, for instance, divides opinion, but not only between Democrats and Republicans. Several rank-and-file Republicans continue to go on the record criticizing and confronting Trump - so there is no reason to believe that this phenomenon does not filter through to ordinary Republican voters on Twitter. In practice, this would mean that people may publish an expressed dislike of Trump on Twitter but still choose to vote for Republican candidates, which would however not be captured by the Trump vote choice classification model.

6.6.1 Implications

There is the question of why the Kavanaugh model performs so well when aggregated to national- and state level vote shares. I suggest that this may be explained by the heavily partisan undertones that developed throughout the evolving story of allegations levelled against him during confirmation hearings for him to become a US Supreme Court judge. It has been widely reported as a major mobilizing factor for supporters of both parties in the midterms (e.g. CNN, 2018a,b; Arkin, 2018) and was the focal point of global news coverage for over a month leading up to election day. Hence, I argue that this model best captures partisan division in how people talked about it on Twitter and therefore produces the best predictions when aggregated.

In terms of the implications for how best to extract signals pertaining to public opinion from Twitter data using this distant supervision method, the Kavanaugh model acts as a useful example outlining fruitful conditions for high-utility estimates. First, I argue that this model works so well in this context as its conversational artefacts can be divided so neatly into two camps - just like the US political system. It is an open question how well this would work in e.g. a multi-party system, or when predicting an outcome variable that is not functionally binary, as in the Republican/Democrat case. However, alternative issue areas with more than two clusters of language patterns could conceivably be identified for such cases. This means that if researchers can define topic areas where conversation is likely to be clustered along partisan divisions, this method of extracting public opinion-relevant signals from tweets is highly appropriate. For instance, this could be applicable for the discussion of salient policy areas. It further seems plausible that widely covered and divisive news stories/current events will most often satisfy these criteria - but topic selection and the resulting keyword definition is clearly a step in the research process deserving of considerable attention.

I argue that this method is likely to perform better with larger sample sizes, and, crucially, larger resulting labelled training sets. Given the resource constraints of a one-person graduate student labelling operation, this paper used a comparatively small samples, which may contribute to low variance in certain relevant aspects. For future research, the rule of thumb should be - the bigger the sample the better. However, this step should only be taken if the same diligence can be applied to analysing the data as in this paper.

6.7 Conclusion

This paper has introduced a new framework for estimating individual-level political preferences from Twitter data without the need for directly observing statements indicating so, or the need for tallying up blunt and noisy mention counts. While the goal was not to forecast elections, the aggregation on the control panel shows that this has the potential of working exceedingly well with a model specified to leverage strong partisan differences in the way users discuss a politically salient topic, in this case the dominant news story of the election cycle. Indeed, this approach, when compared with existing approaches of Twitter-based election forecasts, provides a more robust theoretical framework and replicable method paired with convincing results in the initial pilot.

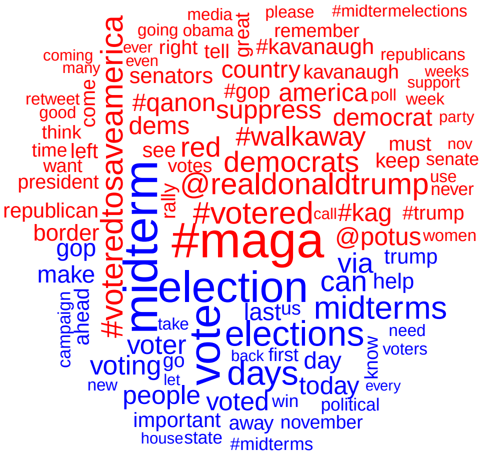
Furthermore, this research has shown the benefits of using randomly selected samples of Twitter users, rather than keyword-filtered collections from Twitter's stream. While a random sample of Twitter users is clearly still a way away from being representative of the offline voting population, it proved a more balanced dataset, both in terms of its aggregate demographics and in terms of more accurately reflecting real-world public opinion.

Future work is required in re-testing this approach in different electoral contexts, such as countries with multi-party systems, as well as new elections in the US. Two separate things should be focused on: one, what is the ideal sample and sampling strategy, and two, what is the ideally specified model that maximises the capture of partisan differences in how people discuss certain topics? However, if one's goal is not to forecast elections, but to classify Twitter users' political preferences on certain politically salient topics, this paper outlines a method which meets precisely those requirements while leaving ample room for adjustments to suit each individual application.

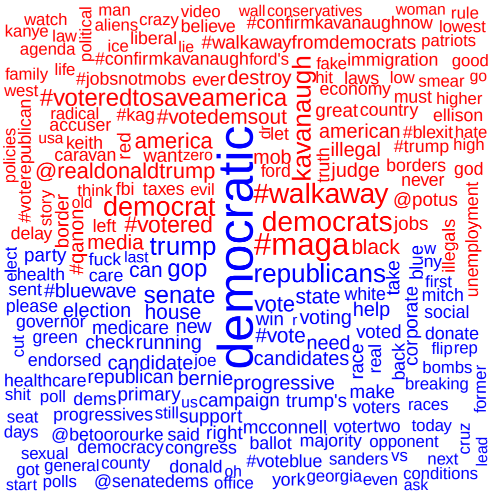
Appendix B

B.1 Wordcloud plots

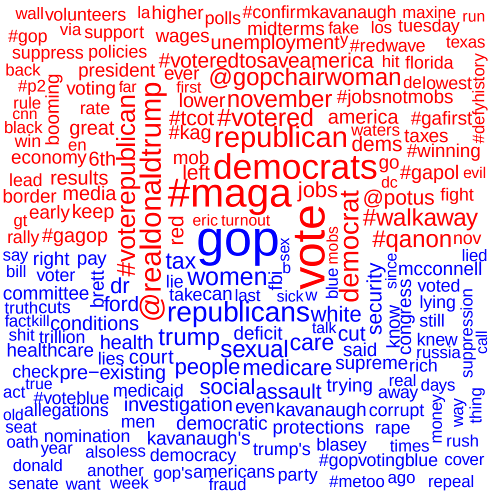
B.1.1 Civic Panel



(a) Midterms model

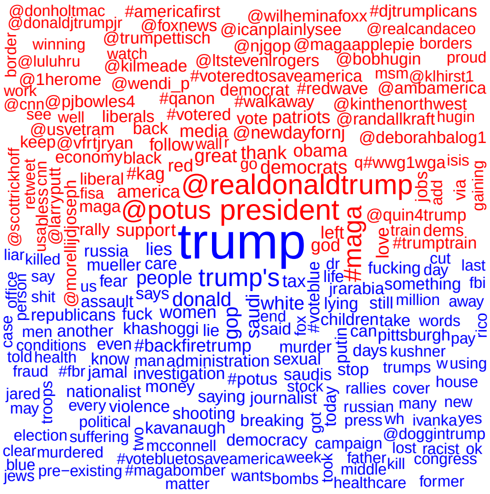


(b) Democrats model

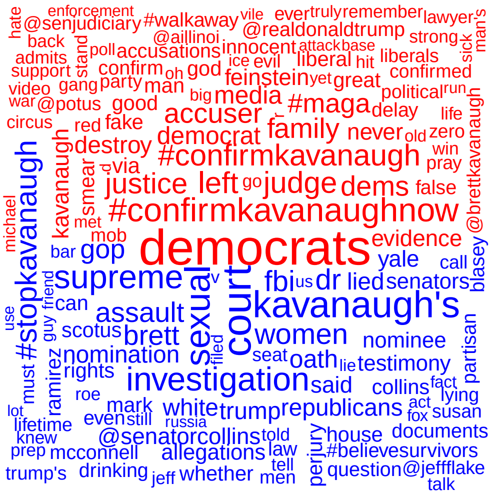


(a) Republicans model

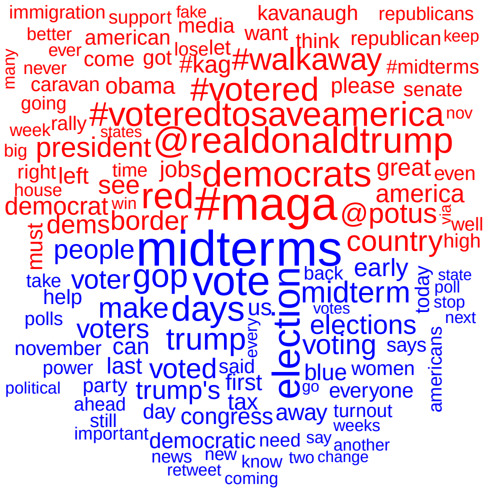
B.1.2 Control Panel



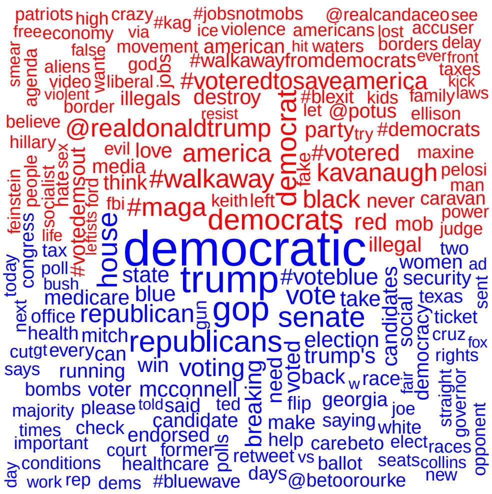
(a) Trump model



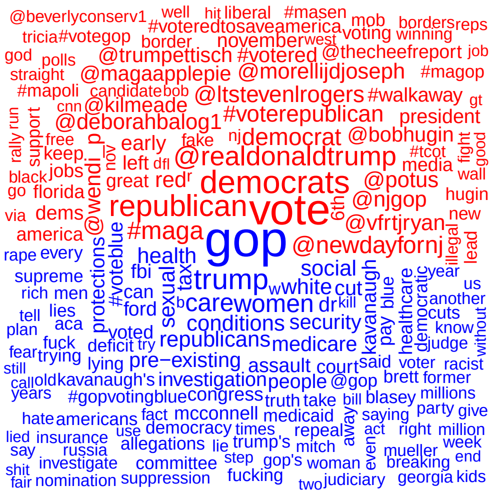
(b) Kavanaugh model



(a) Midterms model



(b) Democrats model



(a) Republicans model

B.2 Logistic regression models

	Dependent variable:		
	V=D	V=R	V=D, SP->CP
	(1)	(2)	(3)
Tweetscores ideology	-1.242^{***}	1.285^{***}	-1.507^{***}
	(0.089)	(0.092)	(0.094)
Sex: male	0.266	-0.434^{*}	0.712^{***}
	(0.249)	(0.256)	(0.252)
Ethnicity: asian	0.535	−0.756	0.890
	(0.853)	(0.862)	(0.840)
Ethnicity: black	1.515^{*}	-1.757^{*}	1.570^{*}
	(0.917)	(0.925)	(0.940)
Ethnicity: latin/hispanic	0.097	−0.305	-0.947^{*}
	(0.566)	(0.571)	(0.555)
Ethnicity: other	−16.694	−0.818	−0.567
	(616.752)	(2.217)	(1.760)
Ethnicity: white	−0.012	−0.394	−0.073
	(0.305)	(0.308)	(0.293)
Constant	0.928^{***}	-0.637^{**}	1.075^{***}
	(0.306)	(0.304)	(0.292)
Observations	607	607	778
Log Likelihood	−241.557	−235.653	−256.048
Akaike Inf. Crit.	499.114	487.307	528.096
Note:			*p<0.1; **p<0.05; ***p<0.01

Table B.1: Logistic regression models predicting classified user-level vote choice (Control Panel)

Chapter 7

Discussion

In the following pages, I discuss the findings presented in this thesis in detail, and further evaluate which lessons can be drawn from them, both in regard to encouraging, successful insights and in regard to less successful, cautionary tales. Furthermore, I relate some of the perhaps less obvious findings to the core questions and debates raised in the literature review, specifically regarding the nature of public opinion and its formation, the relative roles for survey research and digital-trace data based methods in measuring public opinion, and contrasting my approaches - albeit instrumental - to forecasting election results to the ones previously discussed.

7.1 Encouraging findings across all papers

The three paper approach employed in this thesis has yielded findings specific to each of the papers, which have been discussed in detail therein, but crucially, there are certain discoveries which hold true across the three research projects and hence deserve further attention and discussion.

Most importantly, all three applications of my quest towards furthering the understanding of how to measure public opinion using tweets have shown that analyses should be conducted at the user-level rather than the tweet-level. While this may be intuitively plausible, as, how I phrased it previously, it is 'humans who vote, not tweets', the existing body of Twitter-based political science research paints a different picture. This may be due to the fact that the tweet is the form in which data are delivered to the researcher by Twitter's API, and overall the platform is centred much more around the content than e.g. users' profiles. It is nonetheless not consistent with the goals of studying human behaviour to focus on individual records of such behaviour, rather than the humans who it originates from. I have shown the necessity for this in 'Finding Friends', where the user-level top-candidate approach showed a balanced, geographically distinct distribution of preferred presidential candidates, which resembled expectations of where certain candidates might be expected to garner support - the left-wing Bernie Sanders in States with large urban centres and minority populations, the moderate, catch-all Joe Biden in the more southern and suburban, less liberal states, the African American candidate Kamala Harris in the states with high proportions of African Americans, and Mayor Pete Buttigieg in his home state of Indiana, which he never ceased to reference when campaigning. In 'Understanding Political Sentiment', I found that aggregating sentiment annotated measures of political sentiment to state-level vote share estimates in primary elections was significantly more predictive when conducted at the user vs. the tweet-level, while in 'Listening in on the noise', the novel method based on combining *all* available data for a given user offered a way of reliably assigning a likely vote choice to a given user, one which, to my knowledge, forms the most theoretically consistent and empirically accurate (for the Kavanaugh model, considering its consistency across 5 distinct elections) Twitter-based election forecast to date. While there may be research applications in which a tweet-level analysis actually furthers the goals of what is to be studied - for instance, such designs focused on monitoring the emergence of 'trends' on the platform, rather than transposing them to the offline world - I argue that the majority of future Twitter-based social science should focus on the user, not the tweet, and this thesis provides ample amounts of empirical evidence for why that is a good idea.

Besides the unit of analysis, this thesis makes a thorough case for the benefits of the efficacy of attaching Twitter-derived data with computationally inferred estimates for user-level attributes. In this thesis, I have conclusively shown this to be the case both for user-level geography (as outlined methodologically in 'Finding Friends' and applied empirically in all three papers) and user-level political affinity, as implemented using the algorithm developed by Barberá (2015). This is not surprising in principle, as it is no secret that geography and political affinity can account for much of the variance observed in individual-level electorally/politically salient variables. However, there is a difference between conceptually knowing that people will likely vote differently depending on their home state or their partisan affiliation, and computationally inferring quantitative estimates for both from available metadata, and *trusting* that they are valid to the extent that their incorporation moves findings derived from predictive models closer toward the truth they are aiming to measure. Therefore, I am happy that the geo-locating algorithm presented in this thesis will be available to researchers for free, much like the software required to compute Barbera's affinity scores, and will thus, given adoption - and future accessibility of the same Twitter API endpoints - improve and enhance other people's Twitter-based research. Conversely, not all demographic estimation algorithms are equally useful. As shown in "Listening in on the noise", the inclusion of census-derived estimates for user-level ethnicity and sex proved likely too noisy an estimate to improve model accuracy, and indeed, the model in which these data were included performed considerably worse than those that did not, when aiming to accurately model vote share percentages from sentiment-annotated Twitter data. This suggests that much more work is to be done into reliably identifying user-level attributes, and I suggest that the best pathway to a successful implementation likely lies in the expensive and involved approach of using machine learning on users' bios and published content, especially when aiming to infer sex and ethnicity.

Finally, the third core takeaway that holds true across this thesis is one of the importance of sampling for the most accurate possible representations of offline public opinion derived from Twitter. 'Understanding Political Sentiment', with its keyword-derived samples, shows the limitations of such a sampling strategy, while the control panel, a random sample of Twitter users, deployed in 'Listening in on the noise' offers a significantly more balanced representation of what really went on in the target elections. This holds true despite the fact that sample size was considerably lower for the latter than the former. While I can only speculate on this, I suggest that an application of the method used in 'Listening in on the noise' on a random sample of South Carolinian tweeters (if, indeed, such a sample of sufficient size can be constructed) would have resulted in a considerably more accurate vote share percentage estimation - even if many of the tweeters therein might never have mentioned Bernie Sanders, Hillary Clinton, or the primary. Hence, the core takeaway in this regard is that capturing ‘respondents’ (for want of a better term) because they shout ‘here! me!’ is not the best way of sampling, and my research shows this convincingly.

Paper-specific contributions

Besides the overarching themes and takeaways discussed above, there is also a need for highlighting the findings and contributions of the individual papers.

’Finding Friends’ forms an absolute game-changer for the field of Twitter-based social science research. In the future, anyone from an undergraduate student or a big research team can now download this software, geo-locate a set of Twitter users and then analyse their tweets, and draw conclusions on them with considerably more confidence and nuance than before. This is both true for broad social science tasks where a measure of user-level location will add to the validity of the research, and even more so for applications where location is essential. This can now be performed at zero cost, and furthermore, the open source nature of the software package allows others to contribute to its development and maintenance, thereby potentially improving it in the future. While there exist other geo-locating methods, they have all been one-off case studies with inconsistent, un-documented or non-existent replication materials. This package however can be installed, configured and ready to run in 15 minutes with a moderate knowledge of the Linux command line.

A further advancement here is the ‘top candidate’ framing of user-level mentions of presidential candidates. It is considerably closer to candidate support measures one would expect from polls, and hence seems to capture support for a given political candidate much better than its existing benchmark, candidate tweet mention volume.

’Understanding Political Sentiment’ provides a further key contribution in an extension to the overall finding that geo-locating users helps make findings more accurate. Here, I perform intra-state sample stratification by population density, meaning that users from more rural areas, which are less common in Twitter samples are over-sampled in order to more accurately reflect the real-world population density of a given geographic area, such as, in this case, a US state. This improves model accuracy across states and models, and, again, in a future more large-scale application with randomly sampled Twitter users may aid the more widespread adoption of Twitter-based public opinion research.

For ’Listening in on the noise’, the biggest methodological, and indeed conceptual contribution is that this paper does not require users to tweet something directly related to a given concept to be measured (e.g. ’I will vote Republican’) to be included in the analysis, and in order to produce a measure of vote choice for a given user. The fact that I start with a sample of users and then filter this sample downstream to identify *related topics* also increases coverage versus just filtering for keywords directly. In the future, it would be highly interesting to scale this approach and turn it into a large-scale election/public opinion modelling project with significantly larger samples, regular updates of training data, and potentially even a trans-national framework, thereby moving away from the instrumental focus of predicting vote shares in election, and towards a more applied focus of seeking to measure peoples’ opinions on *n*.

The *if*-question: Is Twitter-based public opinion research feasible?

This thesis set out to further investigate the feasibility of Twitter-based public opinion research. As I argued throughout this thesis, the existing and growing literature in this space already provided promising evidence to answer this question in the affirmative prior to the completion of this thesis. So it is no secret that I was cautiously optimistic on this front even at the beginning of this project. However, my goal was to contribute empirical evidence and theoretically and empirically grounded arguments which which cautious optimism may give way to tentative certainty. I argue that evidence from across all three papers provides this. This is achieved as much of the research designs employed in this thesis allow for a comparative assessment of the veracity of findings and the efficacy of applied methods, something that previous publications mostly lack (studies were mostly one-off case studies). Furthermore, the degree to which findings are repeatedly supported throughout the thesis and Twitter-derived measures of public opinion are able to map onto analogous offline measures provides strong arguments in support of the endeavour's feasibility.

In 'Finding Friends', this is most pertinently illustrated by Figure 4.9 (p. 108) which depicts a state-level map of the US, colourised as a reflection of the most frequent 'top candidate', that is, candidate who received >50% of individual users' tweets in the given states. This map a) shows that public opinion measured through Twitter (at least for the case of US Democratic party politics in 2019), while certainly still somewhat biased towards the left when compared to offline measures, *can be adjusted* so as to reflect a more congruent picture of public opinion. Moreover, the adjustment depicted in Figure 4.9 is not an *ad-hoc* approach made to fit data to a known outcome in a way that is agnostic to the theory shaping these adjustments - rather, these data were analysed and adjusted at a point when no single ballot had been cast in the 2020 Democratic primaries, and a map akin to this one would not have looked remotely out of place on the election-relevant reporting of the likes of CNN, Fivethirtyeight or Vox. Furthermore, the adjustments - assessing candidate popularity at the user-level, and setting a high threshold for ascribing a 'top candidate' to a given user - are theoretically consistent and intuitively plausible, making it all the more encouraging that such a data processing step provides a more congruent reproduction of generally accepted patterns of (offline) public opinion. I argue that this is capturing something which is both theoretically consistent and intuitively plausible, given the evidence that exists on US voter preferences, as determined by state-level location and candidate sex and ethnicity.

While 'Understanding Political Sentiment' does not provide findings as intuitively convincing for making a strong, generalisable case for the feasibility of the endeavour of transposing online to offline public opinion, it rather reinforces the *necessity of adjusting/stratifying Twitter datasets* for congruent results. Here again, we see the seemingly baked-in left-wing bias of samples of political tweets and users (or, perhaps, a pro-Bernie Sanders rather than necessarily left-wing bias?), and furthermore clear evidence that adjusting samples based on estimated user-level socio-demographics can improve the external validity of Twitter-reliant designs. Nonetheless, I argue that this paper's contribution is important in considering the overall feasibility of Twitter-based public opinion research, as it compares scenarios and research designs where conditions may be more or less fruitful for Twitter-based research, or where it is wise to make further adjustments, produce large-scale, purposefully generated and stratified samples, and crucially, which steps can be expected to improve the external validity of such research.

'Listening in on the noise' on the other hand provides ample evidence in support of the affirmative answer to the if-question. Not only does it present a theoretically sound approach to measuring individual-level political preferences, it also shows that sampling, i.e. random versus non-random, has a noticeable, again theoretically plausible and consistent effect on results. While the mean absolute error of aggregated vote share predictions may not be *as* low as in the few most accurate previously published Twitter-based election forecasts, I argue that the fact I applied the empirical analysis in 5 distinct electoral scenarios provides a quality of evidence that exceeds that of previous forecasts. I believe that this paper offers a robust starting point for future research in this vein, as it improves upon the baseline, and lessens the need to justify *why* one may want to measure public opinion from tweets, but rather opens up new adjustable parameters and contributions that can further improve practice, and enhance the capabilities and reputation of the field as a whole.

In summary, I argue that the ***if***-question can be considered answered in the affirmative: Twitter-based public opinion research *is* feasible, both in its own right and as an addition to established approaches. Future work should hence concern itself more with how this practice can be improved, how it can be made easier and more efficient, and how it can be made more reliable.

The *how*-question: Methodological takeaways from this thesis

The core contribution of this thesis lies in its methodological innovations, both in terms of innovative workflows, and open source software packages for the estimation of user-level home locations.

First, I find it important to again emphasise the importance of my geo-locating pipeline and open source software package for this entire thesis. Neither 'Understanding Political Sentiment' nor 'Listening in on the noise' would have been conceptually or logistically possible without this software. Of course, previous research has provided multiple approaches to geo-locating Twitter users. But, to my knowledge, none of them shared a software package that can be used to achieve it. And, as I highlighted in 'Finding Friends' - relying solely on GPS-tagged tweets is likely to heavily bias any ensuing analyses (unless, of course, the population of interest is in fact those users who enable GPS-tagging of their tweets). So, I argue that this package, much like the 'tweetscores' package published by Pablo Barbera or the 'urlexpander' package published by Leon Yin and Megan Brown, has the potential of improving the state of the field as a whole. Researchers and developers interested in using this software need not be concerned with the cost of running the software, but only with their ability to implement it, and their available computational capacity, Twitter credentials and data storage.

As I have highlighted throughout this thesis, I argue for increased attention devoted to sampling decisions - from the definition of sampling frames to the eventual analysis of data - and I suggest that the 'Finding Friends' pipeline can significantly help with this endeavour. On the one hand, this can be achieved by simply excluding users not located to a certain desired geographic entity. Furthermore, by importing census-derived population distribution records for a target geographic entity, the population distribution of Twitter data samples can also be weighted in order to more accurately reflect a target population. I present and apply this approach in 'Understanding Political Sentiment', and show that it results in increased external validity in the vote share prediction paradigm. A further strength of this geo-locating pipeline lies in the fact that it can ingest *any* collection of Twitter user ids and output a location estimate for these users, with the further ability to predict the (un)certainty of any individual location estimate. This provides further flexibility for researchers seeking to produce a highly reliable sample of users for a specific geographic entity and where false positives may be of a higher concern than false negatives.

Furthermore, my exploration of how best to extract individual-level political preferences from tweets by hand-annotating them, collapsing them to the user-level, and using machine learning to classify larger user and tweet samples provides ample contribution to best practice in both ’Understanding Political Sentiment’ and ’Listening in on the noise’. I suggest that the distant supervision approach to maximising vote choice (or any political preference of interest) classification provides an improvement over the sentiment analysis paradigm, as it seeks to measure a similar concept, albeit with more focused and apt operationalisations, and, as shown in these papers, potentially better results. I suggest that these methodological innovations, coupled with the opportunities of using geo-locating, provide for ample scope for reliable and insightful Twitter-based public opinion research in the future.

7.2 Open questions and shortcomings

While the core takeaways from this thesis are encouraging, there remain certain issues which this thesis has not addressed, and areas in which questions remain.

First, in general, neither of the two applied papers provide a way of mapping data extracted from Twitter to non-voting or non-attitudes, two aspects of (electoral) public opinion which are nonetheless highly important, and arguably crucial. Obviously, there is more to not voting or not participating in the political process than not tweeting, which is why this, while perhaps easy to implement, would be an overly facile approximation. Clearly, this is an area where surveys have an edge over the analysis of digital trace data, and the only way I can currently conceive of modelling non-voting for twitter users is by surveying Twitter users with this in mind, and then generating machine learning classifiers which can predict non-voting as a function of users’ tweets and metadata. However, this is still far from an ideal mechanism, and does not seem intuitive, easy, affordable or generally transferable from one context to another. However, it is important to note that polling and survey-derived public opinion measures *also* struggle to assess and model these phenomena, in that they are often unable to correctly predict which respondent will turn out to vote and which will not, and further to estimate how this may be distributed in target populations. More attention needs to be given to this in future Twitter-based public opinion research, and while incorporating surveys and machine learning into such an analysis may be fruitful, it is certainly a highly complex, difficult to execute task which will be significantly more expensive than any of the research presented in this thesis.

Second, ’Finding Friends’ and ’Understanding Political Sentiment’ illustrate that Twitter is still enormously biased towards political actors (be they groups or individuals) who are able to garner a degree of support on Twitter which may be disproportionate to their support in the "real world". Of course, I am talking about Bernie Sanders, whose measured support from Twitter as documented in both papers vastly exceeds any measure of support he achieved in opinion polling, or actual election results. It is difficult to conceive of a way in which the vote share percentages for e. g. South Carolina in ’Understanding Political

Sentiment' would *ever* be modelled accurately, given the overwhelming support Bernie Sanders got in the state, or for that matter, most states, as shown in 'Finding Friends' - albeit *on Twitter*, and four years later. It is clear that the demographics of Twitter and the offline, (voting or otherwise) population simply do not line up, and besides heavily involved targeted sampling, it is hard to conceive of a way of dealing with such bias. Such a targeted sampling approach would involve manually 'recruiting' a more balanced set of Twitter users for a given state - and thus essentially be qualitative research - incredibly involved, complex, expensive and time-consuming, with no guarantee that it would actually achieve the stated goals of providing a representative sample. This task would be especially daunting given the ease with which polls - and conventional wisdom - can forecast a state like South Carolina for someone who is *not* a socialist like Bernie Sanders. However, there is an argument to be had that Sanders may have been one of the first politicians with such a devoted and active following on Twitter (after, notably, Obama) and that in the future (and indeed the present and recent past, with examples such as Jeremy Corbyn or Donald Trump), it will be less common to observe such a disconnect between real-world support/attention and Twitter-based support/attention, precisely because a strong, active, *loud* following on Twitter will be necessary in order to be competitive in elections. Consider, for instance, an imaginary electoral match-up between Alexandria Ocasio-Cortez and Pete Buttigieg in a presidential primary in 2028. Who is ahead in South Carolina on Twitter and in the real world may be a lot closer together than it was in 2016.

In summary, the incongruence between offline and Twitter world certainly (still) exists, and entirely adjusting for it/eradicating it presents a more complex task than those approached in this thesis. A fruitful pathway toward addressing this will likely entail an application of my 'Finding Friends' pipeline in order to geo-locate giant samples of randomly selected Twitter users, so as to generate users for any geographic (or otherwise) entity of interest. Then, hand-annotation (the qualitative component) will have to code which users fit certain categories of interest - although some of this may be automatable with new tools for estimating individual-level attributes. Then, new 'civic panels' can be created, in the style of online polling firms like YouGov. Such panels should then identify those users who are atypical for the platform, and seek to over-sample them. A possible approach to addressing the question of identifying non-voting, or a general propensity to vote via Twitter could be a mixed-methods survey with linked social media data approach, which could then use machine learning classifiers to generalise an individual-level probability of voting to larger samples of Twitter users.

Notwithstanding the methodological puzzles this thesis does not provide an answer for, I argue that those which it *does* should be considered a significant step forwards for the field.

Adding to the conversation

Finally, it is useful to address some of the topical threads and debates raised in this thesis' literature review. Specifically, it is interesting to relate back some of the findings and takeaways from this thesis to the discussion of public opinion. I asserted that public opinion was in the past seen as constrained in its breadth and complexity by the 'options' in which individuals might shape their political preferences, provided to people by the mass media. I further posited that perhaps the social media era might erode this gate-keeping role held by the media, and thus broaden the range of expressions public opinion can take. A good reference to this conjectured mechanism is displayed in the wordcloud plots in "Listening in on the noise", which illustrates a giant breadth of diverse viewpoints, both on the left and the right. For instance, Republicans frequently mention ‘QAnon’ in the Trump model. While this term, and the associated subculture/movement may have become a household phrase in 2021, in 2018 it was less so - and yet, it was subsumed in how *some* Republicans talked about Trump, while other Republicans talked about more traditional issues, such as ‘unemployment’, the ‘wall’, ‘jobs’ and ‘America’. The same holds true for Democrats, most of whom were united in talking about ‘Trump’, while some talked about ‘Saudi’, ‘Kashoggi’ or ’Putin’, and others yet talked about ‘Medicare’ and ‘pre-existing’ conditions. While it is unlikely that survey-derived measures of the same phenomenon would yield similar granularity, it is further unlikely that a similar wordcloud, of, perhaps ‘dinner table conversations’ in the pre-Twitter era would have shown the same degree of complexity and diversity in how partisans of opposite stripes construct their political views, and ultimately (at least in a two-party system) their vote. While this is merely indicative of the conjectured erosion of the mass media’s gate-keeping function - indeed, it is still predominantly the mass media who disseminate information about topics as varied as Jamal Kashoggi or pre-existing conditions - it was certainly not the mass media who came up with QAnon, and further linked it to Donald Trump’s presidency.

So, to tie this interesting sub-finding into the conversation about the usefulness of analysing Twitter data versus surveys, it is clear that Twitter-based analyses of public opinion can yield insights which surveys cannot, especially when it comes to providing informational depth and breadth at scale. Indeed, the informational depth generated as a by-product of the method outlined in ’Listening in on the noise’ comes close to that achievable with traditional qualitative methods in the social sciences, such as focus groups, however again with the advantage of the data not being generated by means of a researcher-led intervention, and operating in a completely different dimension when regarding the scale of potential analyses. It would be very exciting to see a strong implementation of such a qualitative-inspired strand of research using Twitter data, which is not limited by the small sample sizes typically associated with it, which could explore the main issues partisans of different stripes find important. This could then in turn add to a future expansion of the research presented in ’Listening in on the noise’.

Given the findings presented in this thesis, I argue that there is a strong case for seeking to combine both digital trace data and survey methods in order to produce maximally reliable and informative findings. Specifically, I suggest that in a case like the one of South Carolina described in ’Understanding Political Sentiment’, or the overall dynamics associated with Bernie Sanders observed throughout the empirical parts of this thesis, survey-derived information - even that which exists in the public domain and is thus accessible without incurring major costs - may be hugely beneficial to these analyses. Conversely, survey research could clearly benefit from the depth of understanding of concepts which shape public opinion provided by a method like that outlined in ’Listening in on the noise’, and further, such a method of eliciting political views from those individuals who may not volunteer them may also prove useful to public opinion researchers availing themselves of surveys, given the current issues associated with response rates and hard-to-reach populations.

A further part of this thesis’ literature review was concerned with the discipline of forecasting, and especially electoral forecasting, outlining how the core takeaways from this area of research may not have been the correctly forecast elections, but rather the enhanced understanding of which factors shape electoral outcomes. I suggest that while the majority of my research was not focused on contributing in this vein, 'Understanding Political Sentiment' highlights an important factor which should be taken into consideration when regarding the determinants of US primary elections. Namely, I found that conversation in states was not centred upon state-specific issues, or even state-relevant information relevant to the election or the campaign, but rather followed the prevailing (social) media narratives of the day[1]. For Massachusetts and South Carolina, this meant that a lot of sampled users were talking about the outcomes of the Iowa and New Hampshire primaries, and celebrating or commiserating their candidate's outcome, rather than expressing their views or voting intentions for their own, upcoming primary. In a sense, then, this might suggest that national-level issues and general horse-race stories play a pronounced role in US presidential primaries.

Overall, I suggest that the framing of the question of how best to measure public opinion from tweets as a problem of election forecasting was a fruitful one. This instrumental approach of seeking to learn as much as possible about a given subject of interest by applying it to a measure a known, highly correlated quantity of interest meant that research could still produce useful and important findings, even if the election forecast itself was not necessarily correct. This is the case for instance in 'Understanding Political Sentiment'. Due to the fact that South Carolina is so badly forecast by the models, I am able to conclude that in order to make Twitter data more fit for purpose in an offline scenario which seems so incongruent from its online counterpart, certain innovations, most importantly better sampling will have to occur. However, this approach has given me the opportunity to identify this as an agenda for future research, thus showing the value of the election forecasting paradigm for Twitter-based public opinion research in the case of this thesis.

Finally, a considerable amount of this thesis' literature review focused on published attempts at measuring public opinion from tweets, and zoned in specifically on the main previously employed methodological approaches to the problem, namely keyword mention volume, and positive sentiment volume in tweets. In the literature review, I ascertained that sentiment-based approaches tend to perform better and hence are likely better suited to the task, but are nonetheless faced with several shortcomings, not least the issue that focusing solely on positively classified tweets removes large amounts of pertinent data from a given analysis. So, how has this thesis contributed in terms of further illuminating which approaches might be best suited for extracting public opinion from tweets? I suggest that, first, this thesis has shown the limits of the sentiment analysis paradigm in Twitter-based public opinion research, and the election forecasting paradigm more specifically: it is not able to overcome the inherent biases in a 'standard' sample of tweets, even (or maybe: especially!) when they are geo-located. This is not changed by the fact that I introduce and trial a new method for moving from sentiment scores to user-level vote choice classifications. Second, I present an alternative method to extracting political views from tweets, as conducted in 'Listening in on the noise'. I argue that this a theoretically more sound, more adaptable method, which, in combination with user-level (rather than keyword-driven) sampling, can significantly improve the utility and external validity of individual-level political views compared to both established methods.

[1] This supports a growing body of evidence highlighting the increasing primacy of national-level issues in US politics (see e.g. Hopkins, 2018)

Implications for academic research using digital trace data

This thesis uses tweets as its form of digital trace data. Tweets were chosen for both the reason that they are available (unlike Facebook data) and that the platform's nature provides an apt arena from which to gather data for the purpose of measuring public opinion. However, I also highlighted that tweets and associated metadata are only *one* type of digital trace data, among a myriad of different such data types, which are created day-in, day-out, and are mostly never seen or even thought of by either their creators or researchers.

Hence the question follows: what then can this thesis add to the understanding of general scientific inquiry using digital trace data, beyond the focus on public opinion? I argue that the focus on the user-level aspect is important for accurate findings, regardless of data type or domain. While an individual-level identifier is not necessarily part of any type of digital trace datum, if it is, it should be employed to learn more about a given individual rather than unique data-generating behaviours in isolation. Furthermore, it is important to note at this stage that user privacy and informed consent are paramount, and this concern should always supersede the benefits which can potentially be reaped from sidestepping it. I believe that I present a strong framework for conducting ethically sound research with individual-level digital trace data in this thesis.

It is also interesting to speculate how the findings and advances presented in this thesis would translate to data gathered from other sources. I suggest that conceptually, much of the methodological approaches outlined in this thesis could also be applied to other data sources. In other words, it will always be valuable to know where Facebook or Reddit users live, or where they can be placed on the spectrum of political ideology. These methods would likely work differently on other data types (if at all), and I suspect that the distribution of values derived from them would also differ from those documented in the world of Twitter. This would be explainable by the particular demographics of Twitter, and conversely the demographics of users of other services and platforms which generate digital trace data, which are also likely not to be a carbon copy of target offline populations. This is especially pronounced when it comes to public opinion research.

Having entered the realm of speculation, one may wonder if there are other kinds of digital trace data which would make the analysis presented in this thesis richer or more accurate? The answer has to unambiguously be yes. Facebook data, even just individual-level metadata without user-level content could strongly enhance this analysis, as its rich offering of information stands in stark contrast to Twitter user-level metadata. Another highly useful data type would be the activity logging data of internet-connected devices, such as smart TVs or smart speakers. And given the myriads of other digital trace data which Twitter users create, or indeed digital trace data which are created by non-Twitter users, in the regard of diversity of data types, the 'more is better' maxim is likely appropriate. However, the question of whether it would be ethically appropriate to incorporate such data into a research project akin to this (if they were indeed available) may be more difficult to answer. Several Facebook users will operate their account in the assumption that it is only visible to those whom they give permission to view it. But, for those users whose Facebook profile is public on the web, it would certainly be ethically appropriate to incorporate their data into such an analysis, if only they were available.

How, then might researchers get around the fact that most digital trace data are not available, if we want to incorporate them into Twitter-based analyses? I argue that future research may seek to develop bespoke software tools allowing individual research participants to grant access to certain elements of their personally generated digital trace data to researchers. While this entails high levels of complexity in both the implementation of an infrastructure and interface for collecting such data, it would allow researchers to side-step the rules governing data access dictated by the owners of digital trace data. This approach would further underline the importance of well-theorised sampling decisions, from target populations and sampling frames to participant recruitment. However, all of this depends on individuals being able and indeed allowed to access the digital trace data generated as a result of *their* actions in the first place. This is not a given. Certain services, like FitBit or Strava require users to pay a subscription fee to access certain types of data which *they* generated, and even that does not automatically mean that individuals are allowed to share them with third parties.

Hence, this thesis indirectly and directly strengthens the case for democratisation of access and governance of digital trace data in the 21st century. This thesis clearly shows that these data carry enormous potential for delivering benefits to humanity and the common good. If Twitter is the only sizeable player providing access to these data, then that potential will however be curtailed considerably. There clearly exist competing risks and benefits when considering the democratisation of digital trace data: on the one hand stands the privacy, agency and self-ownership of individuals, while on the other stands the need for research and innovation to address global problems in an accountable fashion. Nonetheless, these questions will not be addressed if researchers, and indeed individuals do not advocate for a change in how *our* data are governed, and even though it does not form a pertinent part of this thesis - as Twitter data *are* available - I believe this thesis to provide strong arguments for why research using digital trace data in the public interest should be promoted going forward.

7.3 Conclusion

Ultimately, this thesis sought out to investigate the feasibility of Twitter-based public opinion research and election forecasting, as well as contribute methods to improve the state of the art in this field. Given the evidence presented in this thesis, I argue that the answer has to be, yes, it is feasible. The pursuit overall is clearly still in its infancy, and this thesis does not provide a one-stop-shop solution for public opinion measurement which will replace the Gallup company, but: it has shown that public opinion research with tweets, as well as forecasting election results can be fit for purpose - if the right steps are taken to ensure that (a) as much as possible is known about individual users contained in the analysis and (b) the right methods are applied in regard to extracting signals pertaining to public opinion from tweets and transposing them to the offline world, as well as producing a theoretically and empirically appropriate sample. I have identified a number of design decisions which make analyses more accurate and provide two specific methodological innovations which can be incorporated into future research in this area, namely my geo-locating pipeline and the method for estimating individual-level political preferences presented in 'Listening in on the noise'.

It is important to note that, in order for public opinion research relying on digital trace data to be taken more seriously and to achieve a more mainstream appeal, there is ample research that has yet to be conducted. Regarding the specific case of Twitter data, I argue that the most important avenues for future research are threefold.

First, there is an urgent need for further software tools by which user-level attributes can be inferred. In essence, such attributes could be anything used to sort humans into groups, which typically also correlate with politically salient behavioural and attitudinal patterns, namely age, sex/gender, ethnicity/race, occupation, socio-economic status, educational attainment or religious affiliation, to name just a few. I suggest that for some of thes attributes, approaches and heuristics already exist which could be applied to this context: sex/gender, race/ethnicity and to some degree age can be probabilistically classified using methods from computer vision and image recognition, applied on users' profile images. However, when it comes to educational attainment or occupation, it appears likely that researchers will have to avail themselves of user-level content and network features. Clearly, this is not an easy task, but I argue that it is absolutely essential in order to push the field of Twitter-based public opinion research to the next level.

A second major direction for future research needs to be the identification of non-voting and non-participation in Twitter users. While an enhanced understanding of users' demographics will likely aid this endeavour, it forms an even more complex puzzle, as there is no intuitively linked type of metadata or content in Twitter data which might explain this. A tentative, arguably simplistic attempt at addressing this question may avail itself of the research design applied in 'Listening in on the noise', but *invert* it. So, one would first search samples of users for those who explicitly state that they do *not* intend to vote or participate, and then attempt to classify other users in samples based on user-level content on related issues. While it would certainly be insightful to perform such an analysis, I suggest that the explicit publication of non-participation is likely to be far less wide-spread than that of participation, and that it may be a more complex, unique type of behaviour that does not follow the same logic as classifying vote choice through distant supervision. This is why this undertaking may be ideally suited to a hybrid survey and Twitter-based design, whereby respondents are asked to provide their Twitter credentials, or perhaps Twitter users are targeted with a survey. The survey should then contain questions asking whether the respondent intends to vote, whether they voted in the past, and so on. Then, using the respondents' Twitter credentials, it would again be possible to use user-level content and metadata, as well as computed and survey-derived individual-level attributes to classify a probability of non-participation.

Third and finally, I suggest that following the promising findings and methodological innovations presented in this thesis, future research should seek to a) move the focus from election forecasting to one of a substantive interest (besides the area of elections), and furthermore expand the scope of the research to potentially be transnational. Intuitively, I would argue for a similar research design as presented in 'Listening in on the noise', but with significantly larger samples for different countries. Then, the target for measurement could be a topic of substantive concern, such as attitudes on the climate catastrophe or rising inequality. By using the geo-locating method outlined in 'Finding Friends', one would be able to construct reliable large-scale samples and then produce different country or region-specific classifiers which would again be distantly supervised by those in their respective samples who make explicit statements on the topic of interest. Clearly, such a research project would far exceed 'Listening in on the noise', but I suggest that it would have the potential of adding both nuanced empirical insights to the study of public opinion on certain issues, as well as provide a further milestone for measuring public opinion with Twitter data.

Many of the improvements detailed in this thesis, and indeed a large degree of the motivation that led to its commencement, have been concerned with critiques of the state of the field, offered most prominently by Andreas Jungherr or Daniel Gayo-Avello. These authors' influential publications called into question the feasibility of Twitter-based public opinion research election forecasting. In this thesis, I feel I have addressed many of their concerns with the field as a whole: rather than performing one-off case studies without theorising any connection between data and the process that is being modelled, I have presented a thesis which sequentially builds on its own findings, and applies methodological contributions in a more theorised, and most importantly, comparative (that is, multi-case) framework. Furthermore, I have contributed considerably to the previously unaddressed question of how to make Twitter samples more representative by developing a tool by which users' locations can be inferred. So, unlike these authors, I reject the notion of the futility of the pursuit of Twitter-based public opinion research that can be transposed to the offline world, and instead see this thesis as part of an accelerating journey for a discipline still in its infancy. I strongly believe and hope that others will build upon this work, and indeed the work of Jungherr, Gayo-Avello and others, such as Pablo Barbera, to usher in a new chapter of social science and public opinion research in the social media age, with the goal of ensuring that democracies govern with only their citizens' interests as motivation, and the betterment of society as its key motivator.

Chapter 8

Bibliography

Abbasi, A., A. Hassan, and M. Dhar (2014). Benchmarking Twitter Sentiment Analysis Tools. In *LREC*, Volume 14, pp. 26–31.

Agarwal, A., B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, pp. 30–38.

Al Zamal, F., W. Liu, and D. Ruths (2012). Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. *ICWSM 270*(2012).

Allcott, H. and M. Gentzkow (2017, May). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives 31*(2), 211–236.

Alvarez, R. M. (2016, March). *Computational Social Science: Discovery and Prediction*. Cambridge University Press. Google-Books-ID: MqqzCwAAQBAJ.

Anderson, C. J. (2000, June). Economic voting and political context: a comparative perspective. *Electoral Studies 19*(2–3), 151–170.

Arkin, J. (2018). Anger vs. elation: Parties scrap for Kavanaugh edge in midterms.

Armstrong, J. S. (2001, May). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Springer Science & Business Media. Google-Books-ID: ezTaBwAAQBAJ.

Backstrom, L., E. Sun, and C. Marlow (2010). Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pp. 61–70. ACM.

Barberá, P. (2016). Less is more? how demographic sample weights can improve public opinion estimates based on twitter data. Technical report, Working Paper.

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., A. Casas, J. Nagler, P. J. Egan, R. Bonneau, J. T. Jost, and J. A. Tucker (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review 113*(4), 883–901. Publisher: Cambridge University Press.

Barberá, P., J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science 26*(10), 1531–1542.

Barclay, F. P., C. Pichandy, A. Venkat, and S. Sudhakaran (2015). India 2014: facebook 'like'as a predictor of election outcomes. *Asian Journal of Political Science 23*(2), 134–160.

Barr, C. (2016, June). The areas and demographics where the Brexit vote was won. *The Guardian*.

Barros, J. M., J. Duggan, and D. Rebholz-Schuhmann (2018). Disease mentions in airport and hospital geolocations expose dominance of news events for disease concerns. *Journal of biomedical semantics 9*(1), 18.

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

Bekafigo, M. A. and A. McBride (2013). Who tweets about politics? political participation of twitter users during the 2011gubernatorial elections. *Social Science Computer Review 31*(5), 625–643.

Benoit, K. (2018). *quanteda: Quantitative Analysis of Textual Data*. R package version 1.3.4.

Bermingham, A. and A. Smeaton (2011). On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pp. 2–10.

Bollen, J., H. Mao, and X. Zeng (2011a). Twitter mood predicts the stock market. *Journal of computational science 2*(1), 1–8.

Bollen, J., H. Mao, and X. Zeng (2011b). Twitter mood predicts the stock market. *Journal of computational science 2*(1), 1–8.

Bond, R. and S. Messing (2015). Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook. *American Political Science Review 109*(1), 62–78.

Bound, J., C. Brown, and N. Mathiowetz (2001, January). Chapter 59 - Measurement Error in Survey Data. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5, pp. 3705–3843. Elsevier.

Bradburn, N. M., L. J. Rips, and S. K. Shevell (1987, April). Answering autobiographical questions: the impact of memory and inference on surveys. *Science 236*(4798), 157–161.

Brown, M. A., Z. Terechshenko, N. Loynes, T. Paskhalis, and J. Nagler (2020). Debate Twitter.

Buntain, C. and J. Golbeck (2015). This is Your Twitter on Drugs: Any Questions? In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, New York, NY, USA, pp. 777–782. ACM. event-place: Florence, Italy.

Buntain, C., J. Golbeck, and G. LaFree (2015). Powers and problems of integrating social media data with public health and safety. *Bloomberg Data for Good Exchange, New York, NY, USA*.

Burnap, P., R. Gibson, L. Sloan, R. Southern, and M. Williams (2016). 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies 41*, 230–233.

Bélanger, , M. S. Lewis-Beck, and R. Nadeau (2005). A political economy forecast for the 2005 British general election. *The British Journal of Politics & International Relations 7*(2), 191–198.

Caldarelli, G., A. Chessa, F. Pammolli, G. Pompa, M. Puliga, M. Riccaboni, and G. Riotta (2014). A multi-level geographical study of Italian political elections from Twitter data. *PloS one 9*(5), e95809.

Campbell, A., U. o. M. S. R. Center, P. E. Converse, W. E. Miller, and D. E. Stokes (1980, September). *The American Voter*. University of Chicago Press. Google-Books-ID: JeYUrs_GOcMC.

Campbell, J. E. (1991). The presidential surge and its midterm decline in congressional elections, 1868-1988. *The Journal of Politics 53*(2), 477–487.

Center, P. R. (2018, March). Social Media Use 2018: Demographics and Statistics | Pew Research Center.

Ceron, A., L. Curini, and S. M. Iacus (2015). Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the United States and Italy. *Social Science Computer Review 33*(1), 3–20.

Ceron, A., L. Curini, S. M. Iacus, L. Curini, and S. M. Iacus (2016, December). *Politics and Big Data : Nowcasting and Forecasting Elections with Social Media*. Routledge.

Ceron, A., L. Curini, S. M. Iacus, and G. Porro (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society 16*(2), 340–358.

Chandra, S., L. Khan, and F. B. Muhaya (2011, Oct). Estimating twitter user location using social interactions–a content based approach. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp. 838–843.

Cheng, Z., J. Caverlee, and K. Lee (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768. ACM.

Christian, L. M., D. A. Dillman, and J. D. Smyth (2008). The effects of mode and format on answers to scalar questions in telephone and web surveys. *Advances in telephone survey methodology 12*, 250–275.

Clark, T. (2016, May). Phone survey finds 10-point lead for remain but web poll puts leave ahead. *The Guardian*.

CNN (2018a). Republicans needed a midterms miracle. Could Brett Kavanaugh be it?

CNN, A. b. Z. B. W. (2018b). Republicans' midterms secret weapon? Brett Kavanaugh.

Cody, E. M., A. J. Reagan, P. S. Dodds, and C. M. Danforth (2016, August). Public Opinion Polling with Twitter. *arXiv:1608.02024 [physics]*. arXiv: 1608.02024.

Cody, E. M., A. J. Reagan, L. Mitchell, P. S. Dodds, and C. M. Danforth (2015). Climate change sentiment on twitter: An unsolicited public opinion poll. *PloS one 10*(8), e0136092.

Colleoni, E., A. Rozza, and A. Arvidsson (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication 64*(2), 317–332.

Compton, R., D. Jurgens, and D. Allen (2014, October). Geotagging one hundred million Twitter accounts with total variation minimization. In *2014 IEEE International Conference on Big Data (Big Data)*, pp. 393–401.

Conover, M. D., B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer (2011). Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pp. 192–199. IEEE.

Conover, M. D., B. Gonçalves, A. Flammini, and F. Menczer (2012). Partisan asymmetries in online political activity. *EPJ Data Science 1*(1), 6.

Converse, J. M. (2017, July). *Survey Research in the United States: Roots and Emergence 1890-1960*. Routledge. Google-Books-ID: GCAxDwAAQBAJ.

Converse, P. E. (2006). The nature of belief systems in mass publics (1964). *Critical review 18*(1-3), 1–74.

Cooley, C. H. (1918). *Social process*. New York, Scribner's.

Crossley, A. M. (1937, January). Straw Polls in 1936. *Public Opinion Quarterly 1*(1), 24–35.

Culotta, A., N. R. Kumar, and J. Cutler (2015). Predicting the Demographics of Twitter Users from Website Traffic Data. In *AAAI*, pp. 72–78.

Cunha, E., G. Magno, M. A. Gonçalves, C. Cambraia, and V. Almeida (2014). He votes or she votes? Female and male discursive strategies in Twitter political hashtags. *PloS one 9*(1), e87041.

Das, S. and A. Kramer (2013). Self-censorship on Facebook. In *Seventh international AAAI conference on weblogs and social media*.

Dave, K., S. Lawrence, and D. M. Pennock (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pp. 519–528. ACM.

Davis Jr. Clodoveu A., Pappa Gisele L., de Oliveira Diogo Rennó Rocha, and de L. Arcanjo Filipe (2011, November). Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS 15*(6), 735–751.

Davison, W. P. (2017). Public opinion.

Dellinger, A. (2018). Instagram adds new photo descriptions for visually impaired users.

DeVerna, M., F. Pierri, B. Truong, J. Bollenbacher, D. Axelrod, N. Loynes, C. Torres-Lugo, K.-C. Yang, F. Menczer, and J. Bryden (2021). CoVaxxy: A global collection of English Twitter posts about COVID-19 vaccines. *arXiv preprint arXiv:2101.07694*.

DiGrazia, J., K. McKelvey, J. Bollen, and F. Rojas (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one 8*(11), e79449.

Donsbach, W. and M. W. Traugott (2007, December). *The SAGE Handbook of Public Opinion Research*. SAGE. Google-Books-ID: ht5bvEM8FckC.

Duffy, B., K. Smith, G. Terhanian, and J. Bremer (2005, November). Comparing Data from Online and Face-to-face Surveys. *International Journal of Market Research 47*(6), 615–639.

Eady, G., R. Bonneau, J. A. Tucker, and J. Nagler (2019). News sharing on social media: Mapping the ideology of news media content, citizens, and politicians.

Eisenstein, J., B. O'Connor, N. A. Smith, and E. P. Xing (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277–1287. Association for Computational Linguistics.

Erikson, R. S. and C. Wlezien (2012). *The timeline of presidential elections: How campaigns do (and do not) matter*. University of Chicago Press.

Fair, R. (2011, December). *Predicting Presidential Elections and Other Things, Second Edition*. Stanford University Press.

Fair, R. C. (1978). The Effect of Economic Events on Votes for President. *The Review of Economics and Statistics 60*(2), 159–173.

Fairdosi, A. S. and J. C. Rogowski (2015). Candidate race, partisanship, and political participation: when do black candidates increase black turnout? *Political Research Quarterly 68*(2), 337–349.

Franch, F. (2013). (Wisdom of the Crowds) 2: 2010 UK election prediction with social media. *Journal of Information Technology & Politics 10*(1), 57–71.

Freelon, D. and C. Wells (2020). Disinformation as political communication.

Gamon, M., A. Aue, S. Corston-Oliver, and E. Ringger (2005). Pulse: Mining customer opinions from free text. In *international symposium on intelligent data analysis*, pp. 121–132. Springer.

Gayo-Avello, D. (2011). Don't turn social media into another'Literary Digest'poll. *Communications of the ACM 54*(10), 121–128.

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review 31*(6), 649–679.

Gibson, R. K. and I. McAllister (2006). Does cyber-campaigning win votes? online communication in the 2004 australian election. *Journal of Elections, Public Opinion and Parties 16*(3), 243–263.

Giglietto, F. (2012). If Likes Were Votes: An Empirical Study on the 2011 Italian Administrative Elections. In *ICWSM*.

Goel, S., W. Mason, and D. J. Watts (2010). Real and perceived attitude agreement in social networks. *Journal of personality and social psychology 99*(4), 611.

Gomez, B. T. and J. M. Wilson (2001). Political Sophistication and Economic Voting in the American Electorate: A Theory of Heterogeneous Attribution. *American Journal of Political Science 45*(4), 899–914.

Graefe, A. (2013). Issue and leader voting in U.S. presidential elections. *Electoral Studies 32*(4), 644–657.

Graham, M., S. A. Hale, and D. Gaffney (2014). Where in the world are you? geolocation and language identification in twitter. *The Professional Geographer 66*(4), 568–578.

Grant, W. J., B. Moon, and J. Busby Grant (2010). Digital dialogue? australian politicians' use of the social network tool twitter. *Australian Journal of Political Science 45*(4), 579–604.

Groves, R. M. (2011, December). Three Eras of Survey Research. *Public Opinion Quarterly 75*(5), 861–871.

Guess, A., K. Munger, J. Nagler, and J. Tucker (2018, November). How Accurate Are Survey Responses on Social Media and Politics? *Political Communication 0*(0), 1–18.

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

Han, B., P. Cook, and T. Baldwin (2013). A stacking-based approach to twitter user geolocation prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 7–12.

Hancock, J. T., C. Toma, and N. Ellison (2007). The truth about lying in online dating profiles. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 449–452. ACM.

Hansen, L. K., A. Arvidsson, F. Å. Nielsen, E. Colleoni, and M. Etter (2011). Good friends, bad news-affect and virality in twitter. In *Future information technology*, pp. 34–43. Springer.

Harrison, L. (1997). The validity of self-reported drug use in survey research: an overview and critique of research methods. *NIDA Res Monogr 167*, 17–36.

Hecht, B., L. Hong, B. Suh, and E. H. Chi (2011). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 237–246. ACM.

Hillygus, D. S. (2011). The evolution of election polling in the United States. *Public opinion quarterly 75*(5), 962–981.

Hinds, J. and A. N. Joinson (2018). What demographic attributes do our digital footprints reveal? a systematic review. *PloS one 13*(11), e0207112.

Holsti, O. R. (1992). Public opinion and foreign policy: Challenges to the Almond-Lippmann consensus. *International studies quarterly 36*(4), 439–466.

Hong, S. and D. Nadler (2011). Does the early bird move the polls? the use of the social media tool'twitter'by us politicians and its impact on public opinion. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, pp. 182–186.

Hopkins, D. J. (2018). *The increasingly United States: How and why American political behavior nationalized*. University of Chicago Press.

Hopkins, D. J. and G. King (2010). A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science 54*(1), 229–247.

Hox, J. J. and E. D. De Leeuw (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Quality and Quantity 28*(4), 329–344.

Hu, M. and B. Liu (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177. ACM.

Hubbard, B. (2019, November). Why Spy on Twitter? For Saudi Arabia, It's the Town Square. *The New York Times*.

Huettner, A. and P. Subasic (2000). Fuzzy typing for document management. *ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes*, 26–27.

Hyslop, D. R. and G. W. Imbens (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics 19*(4), 475–481.

Iarossi, G. (2006). *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents*. World Bank Publications.

Imai, K. and K. Khanna (2016). Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records. *Political Analysis 24*(02), 263–272.

InternetLiveStats.com (2019). Twitter Usage Statistics - Internet Live Stats.

Johnston, R. J., C. J. Pattie, and J. G. Allsopp (1988). A nation dividing? the electoral map of great britain 1979-1987. In *A nation dividing? The electoral map of Great Britain 1979-1987*. Longman UK/Wiley USA.

Jungherr, A. (2013). Tweets and votes, a special relationship: the 2009 federal election in germany. In *Proceedings of the 2nd workshop on Politics, elections and data*, pp. 5–14. ACM.

Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of Communication 64*(2), 239–259.

Jungherr, A. (2015). Analyzing political communication with digital trace data. *Cham, Switzerland: Springer*.

Jungherr, A. (2017). Normalizing digital trace data. *Digital Discussions: How Big Data Informs Political Communication*.

Jungherr, A., P. Jürgens, and H. Schoen (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social science computer review 30*(2), 229–234.

Jurgens, D. (2013). That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. *ICWSM 13*, 273–282.

Jurgens, D., T. Finethy, J. McCorriston, Y. T. Xu, and D. Ruths (2015). Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. *ICWSM 15*, 188–197.

Kapteyn, A. and J. Y. Ypma (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics 25*(3), 513–551.

Karami, A., L. S. Bennett, and X. He (2018). Mining public opinion about economic issues: Twitter and the us presidential election. *International Journal of Strategic Decision Sciences (IJSDS) 9*(1), 18–28.

Karami, A., A. A. Dahl, G. Turner-McGrievy, H. Kharrazi, and G. Shaw Jr (2018). Characterizing diabetes, diet, exercise, and obesity comments on Twitter. *International Journal of Information Management 38*(1), 1–6.

Katz, E. and P. F. Lazarsfeld (1955). *Personal Influence, The part played by people in the flow of mass communications*. Transaction Publishers.

Kearney, M. W. (2018). *rtweet: Collecting Twitter Data*. R package version 0.6.7.

Keeter, S., K. McGeeney, R. Igielnik, A. Mercer, and N. Mathiowetz (2015, May). From Telephone to the Web: The Challenge of Mode of Interview Effects in Public Opinion Polls.

Kennedy, C. and H. Hartig (2019). Response rates in telephone surveys have resumed their decline.

Kenney, P. J. and T. W. Rice (1983). Popularity and the Vote: The Gubernatorial Case. *American Politics Quarterly 11*(2), 237–241.

Kim, D. S. and J. W. Kim (2014). Public opinion sensing and trend analysis on social media: a study on nuclear power on twitter. *International Journal of Multimedia and Ubiquitous Engineering 9*(11), 373–384.

Kinsella, S., V. Murdock, and N. O'Hare (2011). "I'M Eating a Sandwich in Glasgow": Modeling Locations with Tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, New York, NY, USA, pp. 61–68. ACM.

Klašnja, M., P. Barberá, N. Beauchamp, J. Nagler, and J. Tucker (2017). Measuring public opinion with social media data. In *The Oxford handbook of polling and survey methods*.

Kruikemeier, S. (2014). How political candidates use twitter and the impact on votes. *Computers in human behavior 34*, 131–139.

Lazarsfeld, P. F., B. Berelson, and H. Gaudet (1944). The people's choice.

Lewis, P. (2018, July). 'I was shocked it was so easy': meet the professor who says facial recognition can tell if you're gay. *The Guardian*.

Lewis-Beck, M. S. (1986). Comparative Economic Voting: Britain, France, Germany, Italy. *American Journal of Political Science*, 315–346.

Lewis-Beck, M. S. (1988). Economics and the American voter: Past, present, future. *Political Behavior 10*(1), 5–21.

Lewis-Beck, M. S. (1990). *Economics and Elections: The Major Western Democracies*. University of Michigan Press.

Lewis-Beck, M. S. and T. W. Rice (1982). Presidential popularity and presidential vote. *Public Opinion Quarterly 46*(4), 534–537.

Lewis-Beck, M. S. and T. W. Rice (1984). Forecasting US house elections. *Legislative Studies Quarterly*, 475–486.

Lewis-Beck, M. S. and T. W. Rice (1992). *Forecasting elections*. CQ Press.

Lewis-Beck, M. S. and M. Stegmaier (2000). Economic Determinants of Electoral Outcomes. *Annual Review of Political Science 3*(1), 183–219.

Lewis-Beck, M. S. and M. Stegmaier (2014, April). Weather, Elections, Forecasts: After Richardson. *PS: Political Science & Politics 47*(02), 322–325.

Lewis-Beck, M. S. and M. Stegmaier (2007, August). Economic Models of Voting.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Lin, Y.-R., D. Margolin, B. Keegan, and D. Lazer (2013). Voices of Victory: A Computational Focus Group Framework for Tracking Opinion Shift in Real Time. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, New York, NY, USA, pp. 737–748. ACM. event-place: Rio de Janeiro, Brazil.

Lippmann, W. (1922). *Public Opinion*. Harcourt, Brace. Google-Books-ID: eLobn4WwbLUC.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies 5*(1), 1–167.

Liu, B., M. Hu, and J. Cheng (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pp. 342–351. ACM.

Lui, C., P. T. Metaxas, and E. Mustafaraj (2011). On the predictability of the US elections through search volume activity.

Lupia, P. o. P. S. A., A. Lupia, M. D. McCubbins, and L. Arthur (1998, March). *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge University Press. Google-Books-ID: 2Vv6BhLC6HUC.

Mahmud, J., J. Nichols, and C. Drews (2014, July). Home Location Identification of Twitter Users. *ACM Trans. Intell. Syst. Technol. 5*(3), 47:1–47:21.

Marsden, P. V. and J. D. Wright (2010, April). *Handbook of Survey Research*. Emerald Group Publishing.

Marwick, A. E. and D. Boyd (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society 13*(1), 114–133.

Matsusaka, J. G. and F. Palda (1999). Voter turnout: How much can we explain? *Public choice 98*(3-4), 431–446.

McAllister, I. and D. T. Studlar (1991, August). Bandwagon, Underdog, or Projection? Opinion Polls and Electoral Choice in Britain, 1979–1987. *The Journal of Politics 53*(03), 720–741.

McGee, J., J. A. Caverlee, and Z. Cheng (2011). A geographic study of tie strength in social media. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 2333–2336. ACM.

McKee, S. C. and J. M. Teigen (2009). Probing the reds and blues: Sectionalism and voter location in the 2000 and 2004 us presidential elections. *Political Geography 28*(8), 484–495.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology 27*(1), 415–444.

Mejova, Y., I. Weber, and M. W. Macy (2015). *Twitter: a digital socioscope*. Cambridge University Press.

Metaxas, P. T., E. Mustafaraj, and D. Gayo-Avello (2011). How (not) to predict elections. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 165–171. IEEE.

Miller, A. H. and M. P. Wattenberg (1985, June). Throwing the Rascals Out: Policy and Performance Evaluations of Presidential Candidates, 1952–1980. *American Political Science Review 79*(02), 359–372.

Mislove, A., S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist (2011). Understanding the Demographics of Twitter Users. *ICWSM 11*(5th), 25.

Mocanu, D., A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani (2013). The twitter of babel: Mapping world languages through microblogging platforms. *PloS one 8*(4), e61981.

Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*.

Mosteller, F. (1948). Measuring the error. *The pre-election polls of*, 54–80.

mta.info (2019). mta.info | Developer Resources.

Mueller, J. E. (1970). Presidential popularity from Truman to Johnson. *The American Political Science Review 64*(1), 18–34. Publisher: JSTOR.

Mullen, L. (2018). *gender: Predict Gender from Names Using Historical Data*. R package version 0.5.2.

Munger, K. (2018). Temporal Validity in Online Social Science.

Munger, K., P. Egan, J. Nagler, J. Ronen, and J. A. Tucker (2016). Learning (and Unlearning) from the Media and Political Parties: Evidence from the 2015 UK Election. Technical report, Working Paper.

Nasukawa, T. and J. Yi (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pp. 70–77. ACM.

Newman, B. and A. Forcehimes (2010). "Rally round the flag" events for presidential approval research. *Electoral Studies 29*(1), 144–154. Publisher: Elsevier.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society 97*(4), 558–625.

Nguyen, D., R. Gravel, D. Trieschnigg, and T. Meder (2013). " How Old Do You Think I Am?" A Study of Language and Age in Twitter. In *ICWSM*.

Nooralahzadeh, F., V. Arunachalam, and C.-G. Chiru (2013). 2012 Presidential Elections on Twitter–An Analysis of How the US and French Election were Reflected in Tweets. In *Control Systems and Computer Science (CSCS), 2013 19th International Conference on*, pp. 240–246. IEEE.

O'Connor, B., R. Balasubramanyan, B. R. Routledge, and N. A. Smith (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Icwsm 11*(122-129), 1–2.

Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *International Journal of Public Opinion Quarterly 70*(5), 737–758.

Paik, J. H. and J. Lin (2015). Do multiple listeners to the public twitter sample stream receive the same tweets. In *Proceedings of the SIGIR 2015 Workshop on Temporal, Social and Spatially-Aware Information Access*.

Pak, A. and P. Paroubek (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, Volume 10, pp. 1320–1326. Issue: 2010.

Pang, B. and L. Lee (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, pp. 271. Association for Computational Linguistics.

Pang, B. and L. Lee (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 115–124. Association for Computational Linguistics.

Pang, B., L. Lee, and S. Vaithyanathan (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86. Association for Computational Linguistics.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Popescu, A.-M. and O. Etzioni (2007). Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pp. 9–28. Springer.

Powell, G. B. and G. D. Whitten (1993). A Cross-National Analysis of Economic Voting: Taking Account of the Political Context. *American Journal of Political Science 37*(2), 391–414.

Priedhorsky, R., D. Osthus, A. R. Daughton, K. R. Moran, N. Generous, G. Fairchild, A. Deshpande, and S. Y. Del Valle (2017). Measuring global disease with Wikipedia: Success, failure, and a research agenda. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 1812–1834. ACM.

Punch, K. (2003, April). *Survey Research: The Basics*. SAGE. Google-Books-ID: QGFSr-FVkvvUC.

Purdam, K. and M. Elliot (2015). The changing social science data landscape. *Innovations in Digital Social Research Methods. London: Sage.*

Rahimi, A., T. Cohn, and T. Baldwin (2015). Twitter user geolocation using a unified text and network prediction model. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 630–636. Association for Computational Linguistics.

Rahimi, A., D. Vu, T. Cohn, and T. Baldwin (2015). Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1362–1367. Association for Computational Linguistics.

Rainie, L. (2012, November). Social Media and Voting.

Ram, S., W. Zhang, M. Williams, and Y. Pengetnze (2015). Predicting asthma-related emergency department visits using big data. *IEEE journal of biomedical and health informatics 19*(4), 1216–1223.

Rao, T. and S. Srivastava (2012). Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012)*, pp. 119–123. IEEE Computer Society.

Rattinger, H. and E. Wiegand (2014). Volatility on the rise? attitudinal stability, attitudinal change, and voter volatility. *Voters on the Move or on the Run*, 287–307.

Ren, K., S. Zhang, and H. Lin (2012). Where are you settling down: Geo-locating twitter users based on tweets and social networks. In *Asia Information Retrieval Symposium*, pp. 150–161. Springer.

Rescher, N. (1998, January). *Predicting the Future: An Introduction to the Theory of Forecasting*. SUNY Press. Google-Books-ID: 7za3Z3U2b9EC.

Ribeiro, F. N., M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto (2016). SentiBench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science 5*(1), 23.

Ritterman, J., M. Osborne, and E. Klein (2009). Using prediction markets and Twitter to predict a swine flu pandemic. In *1st international workshop on mining social media*, Volume 9, pp. 9–17. ac. uk/miles/papers/swine09. pdf (accessed 26 August 2015).

Robinson, C. E. (1932). *Straw votes: A study of political prediction*. Ams Pr Inc.

Rodrigues, E., R. Assunção, G. L. Pappa, D. Renno, and W. Meira Jr (2016). Exploring multiple evidence to infer users' location in twitter. *Neurocomputing 171*, 30–38.

Rodrik, D. (2020). Why does globalization fuel populism? economics, culture, and the rise of right-wing populism. *Annual Review of Economics 13*.

Roesslein, J. (2020). Tweepy: Twitter for python! *URL: https://github.com/tweepy/tweepy*.

Sadilek, A., H. Kautz, and J. P. Bigham (2012). Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 723–732. ACM.

Sakaki, T., M. Okazaki, and Y. Matsuo (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pp. 851–860. ACM.

Sang, E. T. K. and J. Bos (2012). Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the workshop on semantic analysis in social media*, pp. 53–60. Association for Computational Linguistics.

Scala, D. J. and K. M. Johnson (2017, July). Political Polarization along the Rural-Urban Continuum? The Geography of the Presidential Vote, 2000–2016. *The ANNALS of the American Academy of Political and Social Science 672*(1), 162–184. Publisher: SAGE Publications Inc.

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics 8*, 341–377.

Severyn, A. and A. Moschitti (2015, August). Twitter Sentiment Analysis with Deep Convolutional Neural Networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, New York, NY, USA, pp. 959–962. Association for Computing Machinery.

Si, J., A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng (2013). Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Volume 2, pp. 24–29.

Signorini, A., A. M. Segre, and P. M. Polgreen (2011). The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one 6*(5), e19467.

Silver, N. (2014, August). Is The Polling Industry In Stasis Or In Crisis?

Simonite, T. (2019). A Health Care Algorithm Offered Less Care to Black Patients. *Wired*.

Sloan, L., J. Morgan, P. Burnap, and M. Williams (2015, March). Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. *PLOS ONE 10*(3), e0115545.

Sloan, L., J. Morgan, W. Housley, M. Williams, A. Edwards, P. Burnap, and O. Rana (2013). Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online 18*(3), 7.

Smith, A. and M. Anderson (2018, March). Social Media Use 2018: Demographics and Statistics | Pew Research Center.

Smith, T. W. (1990, March). The First Straw? A Study of the Origins of Election Polls. *Public Opinion Quarterly 54*(1), 21–36.

spaCy (2020, December). explosion/spaCy. original-date: 2014-07-03T15:15:40Z.

Statista.com (2019). Twitter: number of active users 2010-2018.

Steinert-Threlkeld, Z. (2017). *Twitter as Data*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics 37*(2), 267–307.

Tetlock, P. and D. Gardner (2015, September). *Superforecasting: The Art and Science of Prediction*. Random House.

tfl.gov.uk (2020). Unified API.

Thelwall, M., K. Buckley, and G. Paltoglou (2012). Sentiment strength detection for the social web. *Journal of the Association for Information Science and Technology 63*(1), 163–173.

Theocharis, Y., W. Lowe, J. W. Van Deth, and G. García-Albacete (2015). Using Twitter to mobilize protest action: online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information, Communication & Society 18*(2), 202–220.

Tourangeau, R. (2003). Cognitive aspects of survey measurement and mismeasurement. *International Journal of Public Opinion Research 15*(1), 3–7.

Tourangeau, R. and T. W. Smith (1996, January). ASKING SENSITIVE QUESTIONSTHE IMPACT OF DATA COLLECTION MODE, QUESTION FORMAT, AND QUESTION CONTEXT. *Public Opinion Quarterly 60*(2), 275–304.

Tufte, E. R. (1980). *Political control of the economy*. Princeton University Press.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe (2010). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 0894439310386557.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.

Tönnies, F. (1912). *Gemeinschaft und Gesellschaft*. Google-Books-ID: jGsVAwAAQBAJ.

Vaccari, C. (2013). *Digital politics in Western democracies: a comparative study*. JHU Press.

Varol, O., E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini (2017). Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh international AAAI conference on web and social media*.

Vergeer, M., E. Hermans, and S. Sams (2011). Is the voter only a tweet away? micro-blogging in the 2009 european parliament elections.

Volkova, S. and Y. Bachrach (2015). On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsychology, Behavior, and Social Networking 18*(12), 726–736.

Wang, W., D. Rothschild, S. Goel, and A. Gelman (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting 31*(3), 980–991.

Wang, X., M. S. Gerber, and D. E. Brown (2012). Automatic crime prediction using events extracted from twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pp. 231–238. Springer.

Wasserman, D. and A. Flinn (2019). 2018 House Popular Vote Tracker.

Weng, J. and B.-S. Lee (2011). Event detection in twitter. In *Fifth international AAAI conference on weblogs and social media*.

Williams, T. R. (1959, January). A Critique of Some Assumptions of Social Survey Research. *Public Opinion Quarterly 23*(1), 55–62.

Wojcik, S. and A. Hughes (2019, April). How Twitter Users Compare to the General Public | Pew Research Center.

Wojcik, S., S. Messing, A. Smith, L. Rainie, and P. Hitlin (2018, April). Twitter Bots: An Analysis of the Links Automated Accounts Share | Pew Research Center.

Wyner, G. A. (1980). Response errors in self-reported number of arrests. *Sociological Methods & Research 9*(2), 161–177.

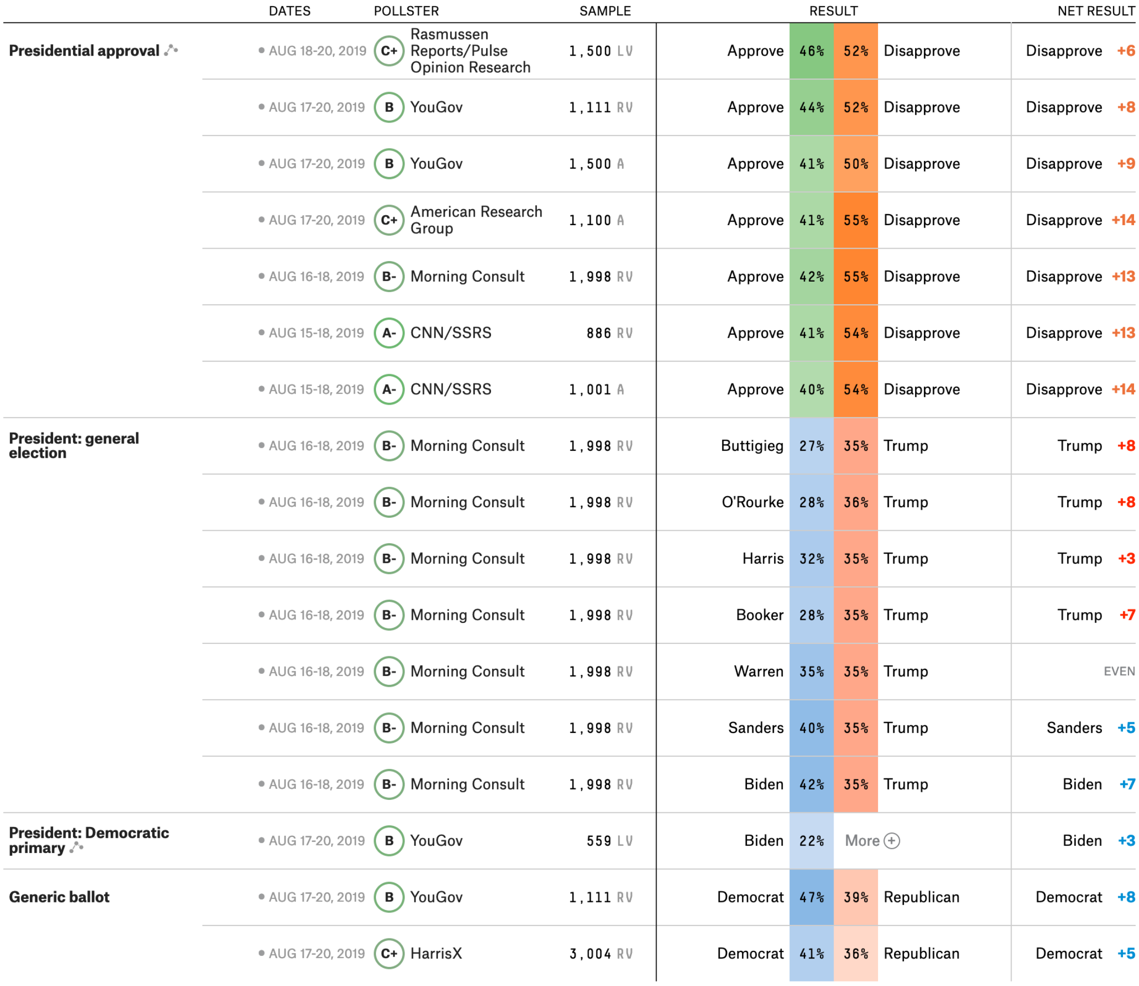
Zaller, J. R. (1992, August). *The Nature and Origins of Mass Opinion*. Cambridge University Press. Google-Books-ID: 83yNzu6toisC.

Zhang, L., R. Ghosh, M. Dekhil, M. Hsu, and B. Liu (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011 89*.

Zhang, X., H. Fuehres, and P. A. Gloor (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences 26*, 55–62.

Appendix C

C.1 Polls from August 21st, 2019 / 538.com poll tracker



	DATES		POLLSTER	SAMPLE		RESULT			NET RESULT	
Presidential approval	AUG 18-20, 2019	C+	Rasmussen Reports/Pulse Opinion Research	1,500 LV	Approve	46%	52%	Disapprove	Disapprove	+6
	AUG 17-20, 2019	B	YouGov	1,111 RV	Approve	44%	52%	Disapprove	Disapprove	+8
	AUG 17-20, 2019	B	YouGov	1,500 A	Approve	41%	50%	Disapprove	Disapprove	+9
	AUG 17-20, 2019	C+	American Research Group	1,100 A	Approve	41%	55%	Disapprove	Disapprove	+14
	AUG 16-18, 2019	B-	Morning Consult	1,998 RV	Approve	42%	55%	Disapprove	Disapprove	+13
	AUG 15-18, 2019	A-	CNN/SSRS	886 RV	Approve	41%	54%	Disapprove	Disapprove	+13
	AUG 15-18, 2019	A-	CNN/SSRS	1,001 A	Approve	40%	54%	Disapprove	Disapprove	+14
President: general election	AUG 16-18, 2019	B-	Morning Consult	1,998 RV	Buttigieg	27%	35%	Trump	Trump	+8
	AUG 16-18, 2019	B-	Morning Consult	1,998 RV	O'Rourke	28%	36%	Trump	Trump	+8
	AUG 16-18, 2019	B-	Morning Consult	1,998 RV	Harris	32%	35%	Trump	Trump	+3
	AUG 16-18, 2019	B-	Morning Consult	1,998 RV	Booker	28%	35%	Trump	Trump	+7
	AUG 16-18, 2019	B-	Morning Consult	1,998 RV	Warren	35%	35%	Trump	EVEN	
	AUG 16-18, 2019	B-	Morning Consult	1,998 RV	Sanders	40%	35%	Trump	Sanders	+5
	AUG 16-18, 2019	B-	Morning Consult	1,998 RV	Biden	42%	35%	Trump	Biden	+7
President: Democratic primary	AUG 17-20, 2019	B	YouGov	559 LV	Biden	22%	More ⊕		Biden	+3
Generic ballot	AUG 17-20, 2019	B	YouGov	1,111 RV	Democrat	47%	39%	Republican	Democrat	+8
	AUG 17-20, 2019	C+	HarrisX	3,004 RV	Democrat	41%	36%	Republican	Democrat	+5

Figure C.1: Screenshot: 538 poll tracker, retrieved 21st August 2019 (*projects.fivethirtyeight.com/polls/*)