MONASH University

Can digital footprints accurately predict political ideology? Evidence from Reddit

Michael Kitchener

Bachelor of Commerce (Honours)

A Thesis Submitted for the Degree of Bachelor of Commerce (Honours) at
Monash University in 2021
Department of Econometrics and Business Statistics

Copyright notice

Abstract

We utilize an original data set from Reddit that contains labels of self-described political ideology for 91,000 Reddit users along with records of the frequency with which they post in over 190,000 different sub-forums (subreddits) and the textual content of their comments to build predictive models that map digital footprints to political ideology. We validate existing findings with our new data set and then leverage our sophisticated response variable, which indicates both the 'social' and economic component of users' political ideology, and allows for centrist self-identification on one or both dimensions, to make several novel contributions to this area. We test how robust predictive accuracy is to the inclusion of centrists and whether economic ideology is more readily predictable from digital footprints than 'social' ideology. Finally, we compare the effectiveness of models trained on records of digital interactions and models trained on the textual content of online comments. We find that the economic dimension of ideology can be predicted to a high degree of accuracy and that this result is robust to the inclusion of those with self-identified centrist economic views. Contrastingly, our models are unable to accurately predict the 'social' dimension of users' ideologies on the basis of digital footprints. We also find that records of digital interactions are stronger predictors of ideology than online comments and discuss why this finding issue may be idiosyncratic to our data. We briefly discuss the policy implications of our findings.

Acknowledgements

Blah blah blah...

Contents

Chapter 1

Introduction

Privacy concerns regarding online data have become increasingly prevalent in the wake of several high profile social media privacy scandals in recent years, most notably the Facebook-Cambridge Analytica data scandal[1]. Of particular concern is the threat that we can inadvertently reveal private traits that we wish to keep hidden through seemingly innocuous online behaviour. This has serious real world implications. Consider the possibility of someone's sexuality being accurately determined from records of their online behaviour. In countries with conservative attitudes towards sexuality, this capability could be abused and people with non-traditional sexual orientations could be at risk of having their sexuality exposed and being subject to prejudice and harm.

Consequently, there is a growing body of academic work devoted to predicting the traits of individuals (personality, gender, sexuality, etc.) on the basis of their digital footprints. Research in this area is important as it illustrates the extent to which this kind of prediction is possible and, therefore, informs the appropriate level of concern that should be paid to this matter. Existing work has largely focused on psychological (e.g. Big Five personality traits) and demographic traits (e.g. gender). For example, it has been shown that neural networks can accurately predict sexual orientation from pictures of an individual[2] and that Facebook Likes can accurately predict Big Five personality traits[3]. Results are best summarized in a 2018 meta-analysis which states "digital traces from social media can be studied to assess and predict theoretically distant psychosocial characteristics with remarkable accuracy"[4].

Political ideology has often been neglected as a focus of this kind of inquiry. However, the extent to which political ideology can be predicted from digital records is particularly salient with respect to the privacy concerns we are interested in exploring. If it is possible to accurately predict an individual's ideology from their digital behaviour (even digital behaviour that is not of an explicitly political nature) then individuals with democratic

sympathies living in authoritarian regimes may be inadvertently revealing their non-conforming political views through ostensibly harmless online behaviour and thus be at risk of harm. Additionally, the possibility of accurately estimating ideology may enable and encourage online political micro-targeting strategies aimed at voter suppression, which is considered to be in opposition with a functional liberal democracy by some[5].

This paper seeks to elucidate the extent to which an individual's political ideology can be determined through records of their online behaviour. We have scraped the usernames and political ideologies of 91,000 users active on Reddit, one of the largest social news and discussion platforms globally. We have also recorded the frequency with which (most of) these users comment and post in a range of interest based discussion sub-forums ('subreddits') as well as the textual content of around 100 comments per user. With this data, we modelled the political ideology of these users as a function of the extent to which they engage with different subreddits and the textual content of their comments.

Our data was scraped from the r/PoliticalCompassMemes subreddit in which users humorously critique ideologies in opposition to their own. In this subreddit, users 'flair' their posts with with their results from the popular Political Compass Test[1]. Figure 1.1 shows an example of how users' comments are flaired with their political ideology. Figure 1.2 shows a post from the subreddit in which a user flaired as a conservative makes fun of U.S. President Joe Biden. Figure 1.3 shows the political compass; users are assigned to one of the four quadrants in the two-dimensional ideological space. These flairs provide an ideological label that we can use as our dependent variable. Chapter 3 provides further information on our data.

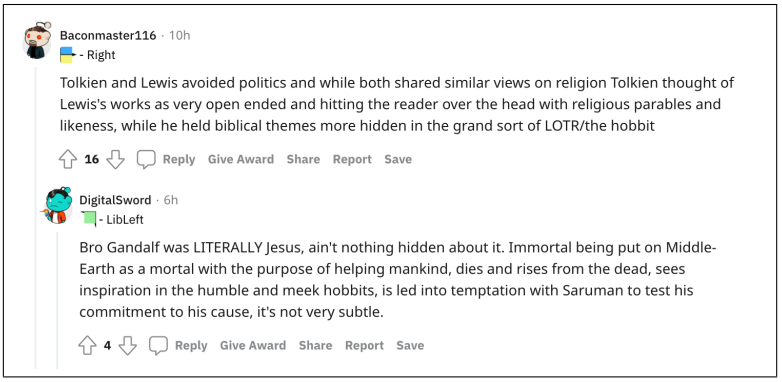

FIGURE 1.1: A comment from a user who has flaired themselves as having economically right wing views and a response from a user who has flaired themselves as a member of the 'libertarian left' (social democrat)

[1]In reality, though users presumably take the test and assign themselves to one of the ideological categories featured in the test, ideological labels should be considered as self-identified ideological placement by users as explained in Chapter 3.

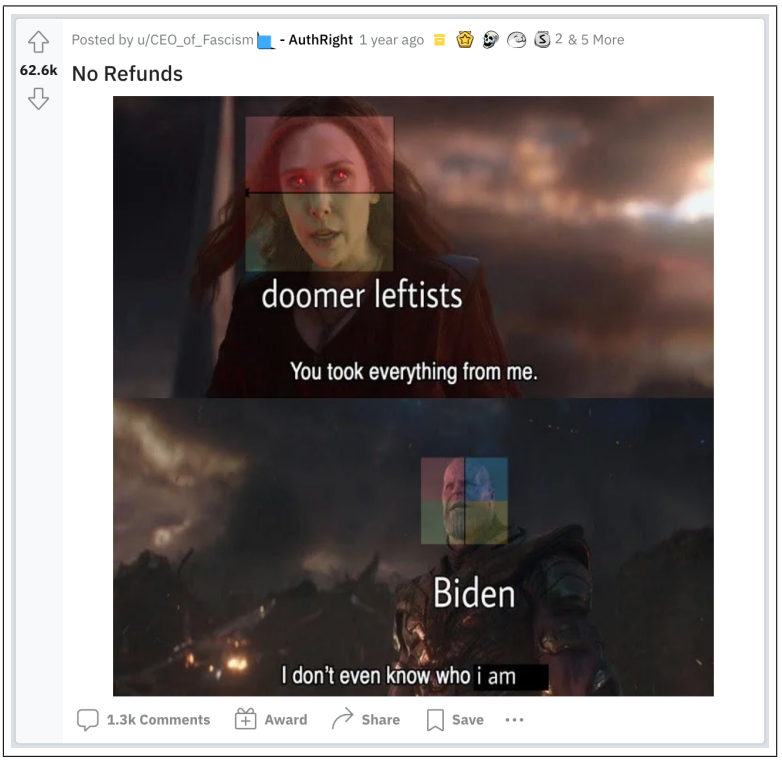

FIGURE 1.2: A post from a user who has flaired themselves as a member of the 'authoritarian right' (traditional conservative)

The possibility of accurately linking online traits is provided for by the fact that an individual's latent psychological makeup drives both their online behaviour and their political beliefs, providing the underlying mechanism for observed association between particular digital behaviors and certain political views. Chapter 2.2 provides a brief discussion of the psychological foundations of ideology. Further, there is a documented association between psychological traits and certain interests[6, 7]. Interests likely impact language use (if you're into movies you probably use words like 'actor' and 'script' a lot) and people with certain interests are more likely to participate in subreddits pertaining to these interests. Psychological traits also drive individual's use of language[8]. This is represented diagrammatically in Figure 1.4.

The upshot of this is that we may be able to model ideology as a function of digital footprints that contain records of a user's interactions with different subreddits or of their language use. In developing these models we aim to contribute to the literature on how, through online behaviour, people may implicitly disclose private information through several avenues. Chapter 4 details the modelling strategies we employ.

Existing findings show that a binary (typically left/right) conception of ideology can be accurately predicted from digital footprints. We investigate whether these results hold for our new, original data set but more importantly we utilize our sophisticated measure of ideology, which allows for centrist self-identification (on both the economic and

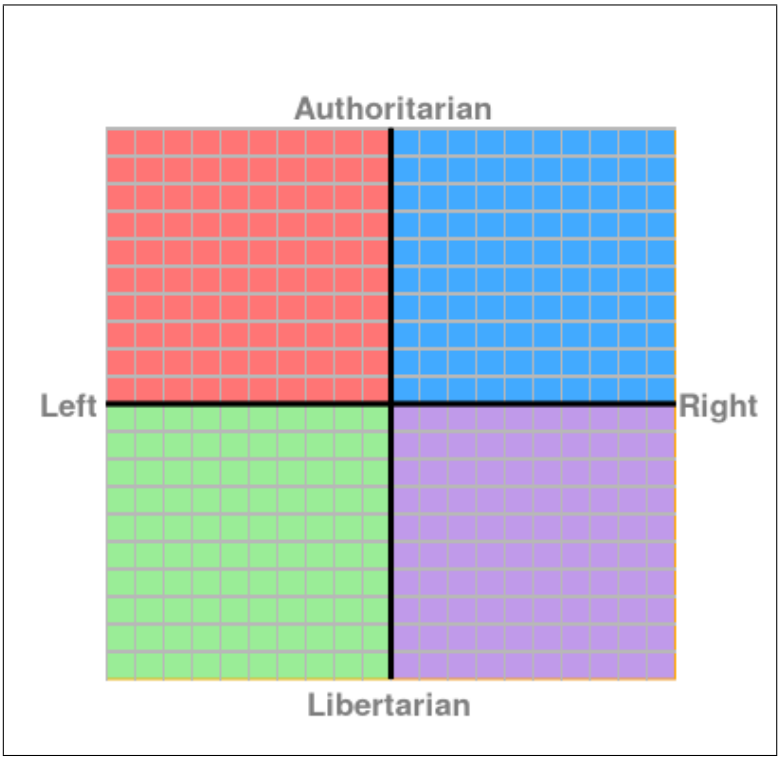

FIGURE 1.3: The political compass. Flaired users take a test specifying the extent to which they agree with politically charged statements (e.g. "If economic globalisation is inevitable, it should primarily serve humanity rather than the interests of transnational corporations") and are assigned to one of four quadrants: libertarian left (social democrat), libertarian right (libertarian) depending on their answers, though some user may identify with just one axis (i.e. purely economically right wing)

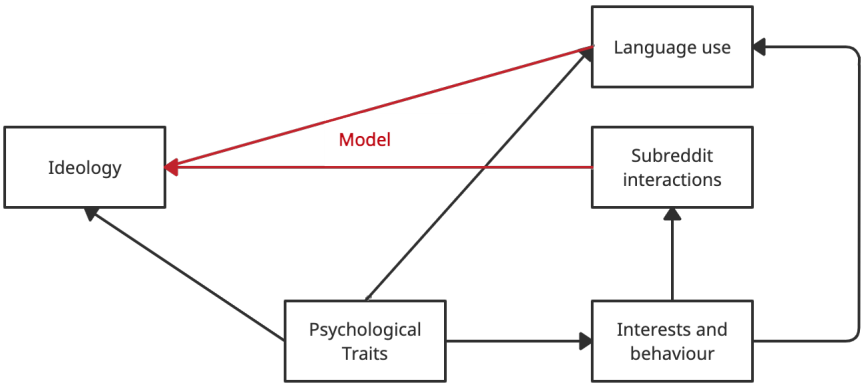

FIGURE 1.4: Causal structure underlying observed association

'social'[2] dimensions of ideology) to test if accurate prediction of ideology from digital foot prints is robust to the inclusion of self described centrists. In a similar vein, we also aim to illustrate how the quality of predictions is affected by the complexity of the ideological variable we seek to model. Further, our sophisticated response variable allows us to decompose users' ideology into both an economic component and 'social' component. This allows us to investigate whether the economic component of someone's political ideology is more readily predictable than the 'social' component. To our knowledge, no other study in this field has attempted to answer these questions. Finally, we model ideology using as features the frequency of users' interactions with subreddits, the textual content of their texts and the union of these sets of features. This provides insight into what sorts of digital footprints are most informative with respect to ideology. These results have important policy upshots that we discuss in Chapter 6.

We also aim to illustrate the usefulness of Reddit as a source of 'social data' to other scholars in this field. The vast majority of studies in this field rely on Twitter and Facebook data. Facebook data is no longer readily available in response to increased public scrutiny over online privacy. Twitter, though a useful source of digital data, only contains textual information. By showing Reddit to be a reliable source of rich social data we hope to spur further developments in this space.

We find that economic ideology is readily predictable from digital footprints and that this result is robust to the inclusion of self-identified economic centrists, although model performance diminishes significantly. Contrastingly, our models are unable to predict 'social' ideology at levels significantly above performance baselines.

We also show that classification of a complex dependent variable representing nine different ideological classes from digital footprints can be done with accuracy above baseline levels. In all cases, the frequency of interactions with different subreddits (analogous to digital footprints like search logs, purchase histories, etc.) is a stronger predictor than the textual content of comments and the union of both sets of features only marginally improves performance.

We provide a fuller account of our results in Chapter 5. 5.1.

[2]The term 'social' for this ideological dimension is somewhat misleading, we provide a fuller explanation of what is meant here in Chapter 3.

Chapter 2

Key Literature

This chapter will lay out some of the key results in the broader area of 'social data science' focused on exploring the possibility of predicting private traits from digital data. It will also lay out the psychological basis of ideology and attempts at modeling ideology as a function of psychological traits.

2.1 Social data science

There is a growing body of work illustrating how online behaviour implicitly discloses private traits. Pioneers in linking traits to digital behaviour relied on volunteers granting access to their social media information and were thus constrained to relatively small sample sizes [9–12]. More recently, researchers have begun using larger data sets with thousands of users or posts as the basis of analysis to make more robust findings. It has been shown that sexual orientation can be accurately predicted from pictures of an individual's face with neural networks[2] and that Facebook Likes can accurately predict Big Five personality traits, gender, race, drug use and other traits[3]. Even high level features of a Facebook profile (number of friends, number of statuses posted) have been shown to correlate with and predict personality[13].

Indeed, it has been found that digital footprints can be used to train statistical models capable of predicting people's personality traits to a higher degree of accuracy than their close friends and family[14]. Publicly available information from Twitter profiles[15, 16] and Facebook messages[8] has also been used to predict these traits using natural language processing techniques. Meta analyses claim that "the predictive power of digital footprints over personality traits is in line with the standard "correlational upper-limit" for behavior to predict personality"[17] and that psychosocial traits more broadly can be accurately predicted by digital footprints[4].

Further, it is not only digital traces from social media that can be effectively employed to predict private traits. The word choices of blog writers[18], the images 'favorited' by Flickr users[19] and mobile phone usage data[20] have also be used to predict personality traits.

These sorts of models can be effectively used to persuade people; S.C. Matz et al., have shown, using Facebook advertising with different messages aimed at people with different 'Likes' indicating particular personality types, that engagement with advertised content can be effectively increased through tailoring messages to predicted psychological traits[21].

Some attention has been payed to political traits too. Natural language processing techniques can be used to accurately classify the political leanings of Twitter users[22, 23] and their political engagement or ideological extremity[24]. Hashtags and network analysis of mention/retweet networks can also be leveraged to effectively classify the ideologies of Twitter users[23]. Kosinski et. al's seminal work[3] also pays some attention to political traits. Kosinski et al. map the Likes of 9,572 American Facebook users with a self-disclosed party preference of either Democrat or Republican using a logistic regression model, illustrating that the model's predictions on unseen data achieve an ROC-AUC of 0.85. Our approach is based off of this paper and aims to improve upon it in several ways detailed in Section 3.5 and through the use of a broader range of models expounded in Chapter 4.

2.2 Psychological basis of ideology and psychological modeling of ideology

Literature in psychology indicates that political views are, in part, determined by psychological traits. We believe this provides for the observed association between online behaviour that is not of an explicitly political nature (commenting on the r/Marxism or r/Conservative subreddits would constitute explicitly political online behaviour) and ideological views.

Much of the work here has focused on modeling ideology as a function of scores on particular psychometric scales. In one study, the Big Five model of personality was used to predict party preferences in Italy, Spain, Germany, Greece and Poland[25]. It was found that they were more powerful predictors than demographic traits (age, gender, income, education) and that openness to experience predicts left wing views whilst conscientiousness predicts voting for conservative parties. They note that the linkage between

personality and political views is consistent with theoretical expectations: conscientiousness comprises a preference for order and adherence with rules whereas openness to new ideas correlates against preferences for rigid social hierarchy[25]. It has been shown that including scores from the HEXACO model of personality as additional regressors alongside Big Five traits, increases the proportion of variance explained when modeling ideology[26].

Another study illustrates that, in addition to openness predicting left-wing views, Hartmann's Boundaries Questionnaire can predict conservatism[27]. Some results are consistent with theoretical expectations; conservatism is correlated with a preference for rigid and well defined social structures whilst other results are harder to explain; conservatism is also positively correlated with a preference for blurred edges and soft lines in art[27]. This suggests the psychological underpinnings of ideological belief may go further than expected.

More broadly, a meta-analysis of 'conservatism as a function of social cognition' has found that death anxiety, system instability, dogmatism, openness to experience, uncertainty tolerance, need for order, structure and closure, integrative complexity, fear of threat and loss and self-esteem all have predictive power of conservatism[28]. Consider how an individual's interests and activities are likely influenced by these traits. Someone with substantial death anxiety may resultingly have a strong interest in existential philosophy. Someone with low openness to experience may prefer older, familiar television shows to more recent ones.

Consequently, psychological traits influence both ideology and our interests, which in turn influences our online behaviour, providing for an observable mapping of online behaviour to ideology which can be approximated by statistical learning models.

Chapter 3

Data

To understand the extent to which digital footprints can be used to predict ideology, and thus elucidate the relevant privacy concerns, we require digital footprints from a large number of internet users whose ideologies are known. This is not a trivial task. The MyPersonality data set that contains information on a large number of Facebook users' Facebook Likes is no longer available and obtaining new data from Facebook is not realistic due to Facebook's increased protection of people's data in response to privacy scandals. Twitter, whilst a viable source of data for such inquiry is limited by the fact that there is no real way to effectively and easily collect information on the ideologies of large amounts of users. Some of the studies mentioned in Section 2.1 required users to voluntarily report their personality types (which was the dependent variable in these studies) and thus had sample sizes in the hundreds or low thousands[15, 16, 23] which is a significant limitation. Others must rely on heuristic and easily fallible methods to obtain ideological labels, such as assigning users to different ideologies or political parties on the basis of which users they follow on Twitter[22] or human annotation of ideology[23].

Further, using purely textual data (like Tweet data) will allow us to predict ideology as a function of explicit textual communication. However, we would also like to get an indication of how more subtle digital footprints that do not involve explicit communication can be used to predict ideology.

As such, we opted to scrape an original data from the r/PoliticalCompassMemes subreddit where hundreds of thousands of users have 'flaired' their posts with their results from the Political Compass test. We created a Python script using the Python Reddit API Wrapper (PRAW) library to scrape the r/PoliticalCompassMemes subreddit and collect the usernames and ideologies of 91,000 users who had commented on posts in the subreddit. We also created a script using the PMAW pushshift.io API wrapper to

collect a maximum 100 comments from each user. Consequently, there were three steps to collating the data:

1. Creating a list of usernames and the ideology signaling flairs associated with each username.
2. For each user in (1), going through their post and comment history and recording a tally of the amount of times they have posted and commented in each subreddit they interact with.
3. For each user in (1), collecting the textual content of the 100 most recent comments they had made in subreddits excluding r/PoliticalCompassMemes.

Kosinski et al. note that findings from studies using Facebook Likes likely generalise to other types of digital footprints[3]. The same is true of Reddit interactions. Reddit users are able to interact with millions of subreddits, which are interest or topic based discussion forums. As such, interaction with a subreddit can be seen as an expression of interest in the relevant topic (if a user interacts with the r/gaming subreddit, they are 'showing' an interest in video games). Consequently, the frequency with which someone interacts with different subreddits is broadly analogous to other (likely richer) digital footprints such as web search logs (searching a topic also indicates interest in it) so our results should generalise to the kinds of digital footprints that entities which threaten digital privacy have access to.

3.1 Creating a list of usernames and ideology signaling flairs

The process of creating a list of usernames and the associated ideology signaling flairs was achieved through running a Python script that cycled through the top 1000 most popular posts of all time in the r/PoliticalCompassMemes/ subreddit. For each post, we looped through all the available comments. If the author of the comment was not already in a list of users whose ideology we had recorded and their comment was flaired with an ideology, we added their username and their flair as a row to the data set. The username was also added to the list of users whose ideology we had recorded to avoid doubling up.

This resulted in a data set of 91,000 username and flair combinations. An illustration of this data can be found in Table 3.1.

TABLE 3.1: Example user-flair data

username	ideology
user1	Libertarian-Left
user2	Authoritarian-Right
user3	Libertarian-Right
...	...

This data is central to the creation of a predictive model that can classify the ideology of Reddit users based on their digital footprints. The flairs signal the ideology of users which is the response variable we are aiming to predict.

There are a total of 9 classes (different flairs that correspond to the same ideology were grouped together as detailed in Appendix B):

- **libleft:** social democrats, i.e. supporters of Bernie Sanders.
- **libright:** libertarians, i.e. supporters of Ron Paul.
- **libcenter:** those who identify as libertarians with respect to the role of the state, but with centrist economic views.
- **centrist:** those who identify as having centrist views on the economy and the role of the state.
- **left:** those who identify as having left wing economic views but centrist views on the role of the state.
- **right:** those who identify as having right wing economic views but centrist views on the role of the state.
- **authright:** traditional conservatives, i.e. supporters of Ted Cruz.
- **authleft:** communists, i.e. supporters of Xi Jinping.
- **authcenter:** those who identify as having authoritarian views with respect to the role of the state on social issues but centrist (or perhaps mixed) economic views[1].

The frequency and sample proportion of each ideological class are described in Table 3.2 and illustrated in Figure 3.1.

The Political Compass Test only allocates users to one of four quadrants (authleft, authright, libleft, libright) by providing quantitative scores of how right or left they are economically and how libertarian or authoritarian they are 'socially'. Presumably,

[1]This ideology is treated as a pro-fascist ideology in the forum.

TABLE 3.2: Ideology frequency and proportion in sample (n = 91,000)

	libleft	libright	libcenter	centrist	left	right	authright	authcenter	authleft
Frequency	18,070	15,054	13,548	13,408	9,646	6,526	5,672	4,801	4,275
Proportion	0.2	0.17	0.15	0.15	0.11	0.07	0.06	0.05	0.05

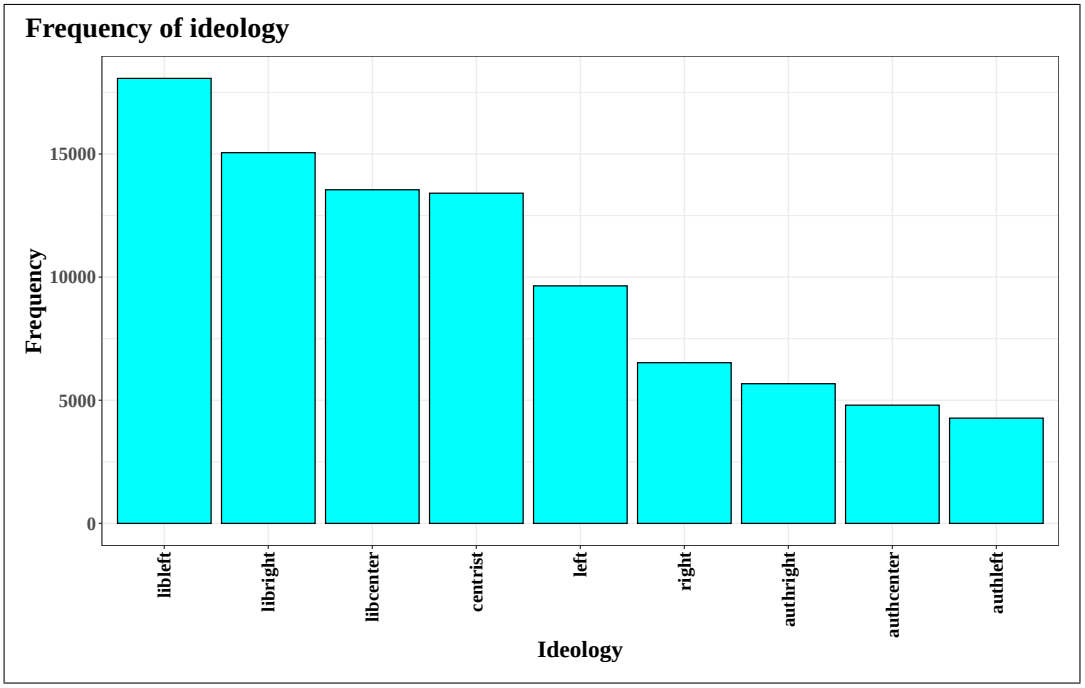

FIGURE 3.1: Ideology frequency in sample ($n = 91,000$), note: we were unable to gather footprints for some users to the size of the data set that is used in model training varies depending on the particular model

users who identify as, say, 'left' have taken the test and found that though they score highly 'left' in the economic dimension but are only marginally above or below the origin (midpoint) of the 'social' axis. Likewise, centrist users likely consider themselves insufficiently far from the origin in either dimension to warrant any other classification. Still, it is fundamentally up to the discretion of the user as to whether to identify in accordance with their test outcome or as a centrist on one or both ideological dimensions. This choice may be based on prior belief regarding their political views.

Consequently, the labels here must fundamentally be understood as self-reported ideologies rather than the results of the test even though self-reported ideologies are presumably strongly influenced by test results.

In what follows we refer to position on the left/right axis as economic ideology (position here is defined by responses to questions regarding the economy) and position on the authoritarian/libertarian axis as social ideology. It may not be immediately clear how the terms authoritarian/libertarian relate to 'social' political ideology but inspection of the relevant questions on the Political Compass Test reveals that these questions elicit preferences for government's coercive involvement in social facets of life, i.e. whether

the government should enforce a ban of abortion. As such the term 'social' ideology provides a useful shorthand that will be employed in the remainder of this paper (with the omission of single quotation marks).

There is no guarantee that each username in our data set corresponds to a unique person; one individual could have two or more different Reddit accounts. However, this is unlikely to be a prevalent behaviour and should have a negligible effect on the validity of our data. It is also possible that an account's flair may not truly represent the ideology of the person behind the account. This could be due to users mistakenly assuming that they subscribe to a particular ideology without having taken the test. It is also possible that some users create accounts flaired with opposing ideologies in order to post unpopular views or stereotypes associated with that ideology in an effort to satirize the beliefs of their ideological opponents. We do not consider this a substantive issue in our data.

It would be possible to obtain a larger data set through several workarounds (looping through different sets of posts and combining the results) and could be done in the future to expand upon this work but the data set is large enough to represent a substantive step forward from other work done in this area including Kosinski's modelling of ideology with Facebook Likes[3].

3.2 Collecting the subreddit interaction records of flaired users

With the user-flair data obtained, the next step was to obtain comprehensive digital footprints for all of the users whose ideology we know. Comprehensive digital footprints can be converted to set of predictor variables for each user in a number of ways, allowing us to model ideology (as signaled by flair) as a function of digital footprint. For each user, we loop through the 1000 most recent posts and 1000 most recent comments. Each post or comment was stored as a row in a data set with the following columns:

- **User:** the username of the user whose posts/comments we are recording.
- **Interaction:** whether the interaction being recorded is a comment or a post.
- **Title:** the title of the post if the interaction is a post, otherwise blank.
- **Score:** the 'score' of the post/comment; other users can 'upvote' a post or comment and increase its score by 1 or 'downvote' the post or comment and decrease its score by 1.

- **Time:** the time the post/comment was made.
- **Subreddit:** the subreddit the post/comment was made in.

A hypothetical illustration of this data can be found in Table 3.3.

TABLE 3.3: Example user-history data

username	interaction	title	score	time	subreddit
user1	post	GTA V ...	43	10:32 2/1/21	r/gaming
user1	comment		12	16:12 8/12/20	r/classicalmusic
user2	post	Today's ...	-6	5:36 4/3/21	r/boxing
user2	post	The ...	3	4:24 4/3/21	r/seinfeld
user2	comment		0	5:09 3/3/21	r/crypto
...	...	...	...	...	...

This can be transformed into a user-interaction matrix where each row represents a unique user and each column refers to a particular subreddit. The value in any particular cell can represent a number of things: the number of times the user has posted/commented in that subreddit, whether the user has posted/commented in that subreddit or the average score of the user's posts/comments in that subreddit. Table 3.4 illustrates what a user-interaction matrix may look like.

TABLE 3.4: Example user-interaction matrix

username	r/gaming	r/classicalmusic	r/boxing	r/seinfeld
user1	3	12	0	0
user2	0	4	1	6
user3	0	0	0	0
user4	1	0	0	14
user5	0	43	0	0
...	...	...	...	...

The user-interaction matrtix is a set of predictors for each user and can be merged with the user-flair data. This allows us to model ideology as a function of digital interactions. In order to examine the extent to which ideology may inadvertently be revealed we removed columns that represent interactions with explicitly political subreddits. Since there are a huge number of (debatably) political subreddits we are unable to remove all of them, however, by removing the many of the most popular political subreddits we hope to minimize the influence of explicitly political subreddits on our models.

The user-history data set is very large (63,709,041 rows) so transforming it into a user-interaction matrix is not a trivial computational task. The pivoting process had to be done in chunks, with the final user-interaction matrix comprising the union of all chunks. We transformed the user-history data set into a user-interaction matrix where

each cell represents the amount of times the user has posted or commented in the relevant subreddit. In the model training step we often map the original user-interaction matrix into a simplified form where each cell represents whether the relevant user has posted or commented in the relevant subreddit at all.

It should be noted that the limit of 1000 posts/comments is not arbitrary. PRAW limits requests to 1000 objects at a given time; i.e. if we are looping through a particular subreddit's top posts we can at most request 1000 post objects from PRAW. It was on this basis that we elected to gather usernames and flairs through the top posts since these are likely to have many comments and hence more user/flair combinations to record. We also chose to record each user's 1000 most recent posts and comments as these are presumably most reflective of the users' most current attitudes and interests. It should also be noted that there may not be exactly 1000 records returned from any request since deleted posts may be returned and contribute to the request limit despite being of no use to us. This issue is discussed in Section 3.4.

If our user-history script encountered an error whilst collecting records of subreddit interactions for a given user (this could happen occasionally due to issues with Reddit's servers or user's settings) it disregarded all data collected for that user and skipped to the next user to avoid censoring records of interaction for specific users.

3.3 Collecting the comments of flaired users

We looped through each user in our list and recorded their 100 most recent comments from any subreddit excluding r/PoliticalCompassMemes. The process of extracting features to use in our models from this raw textual data is described in Section 4.2.

The *comment-manipulator.py* script concatenates all comments from each user and cleans the text, generating two different data sets from which we extract tf-idf and average Word2Vec features.

In the user-interaction matrix, we removed interactions with popular, explicitly political subreddits from our set of features.

No such actions have been taken for the textual comments collected for each user. We do not remove explicitly political words (i.e. words that may be associated with partisanship) as there is no way to a priori determine how the use of politically charged words discloses ideology. For instance, a liberal may use the word 'Trump' and 'gun' a lot as they criticize the Trump administration and lack of stringent gun control in the

United States. That is to say the use of these politically charged words does not explicitly disclose an ideological viewpoint (as we assume participation in explicitly ideological subreddits to). Consequently, we do not remove any words from our corpus.

We could have avoided scraping comments from our list of politically explicit subreddits but there is no reason to think that the language used in these subreddits constitutes explicit disclosure of ideology. If a user comments in the r/conservative subreddit, there is no reason to think that the actual textual content of their comment reveals their ideology; knowing the textual content of a comment in an explicitly political subreddit without knowing what the subreddit actually is, is insufficient information to infer ideology. Perhaps users reserve aspects of language use for explicitly political communities (i.e. racially charged slurs) but the argument that this use of language constitutes an explicit disclosure of ideology in the same way that the act of posting in an explicitly political subreddit does is tenuous.

The scraper occasionally encountered errors when collecting comment data. When this happened, we used an error catch to drop all data for the user whose data we were collecting when the error occurred (to avoid censoring the textual content of comments for specific users) and skipped to record the comments of the next user in our list.

3.4 Caveats

As with all digital data, there are some caveats that should be mentioned. As discussed in the previous section, our digital records of comments for a given user are limited to the maximum of all the user's comments or the 1000 most recent of the user's comments (or possibly less if some of their comments have been deleted or removed). The same applies for our records of users' posts. We also only record the textual content of a maximum of 100 of their comments. This is endemic to data scraped from the web where API limits often restrict how much data can feasibly be extracted for each user. This issue features in related works [15, 24].

Further, our sample is not representative of any particular national population. Users may come from any country (although most appear to come from English speaking countries). The distribution of key demographic variables like age, education, income and race in the population of Reddit users is likely far less varied than it is in any country level population. Further, users who comment and flair themselves in the r/PoliticalCompassMemes subreddit likely have different traits to the broader population of all Reddit users. There may also be systematic differences between the minority of users with privacy profiles that were not amenable to our scraping techniques and those that

were. Clearly there are major sample selection issues here. We are not attempting causal inference but there is a risk that the average r/PoliticalCompassMemes commentor has a greater interest in politics than the average Reddit user. This may manifest in greater engagement with activities (and hence subreddits) which are strongly associated with particular ideologies as ideology is a larger part of personal identity for these people. As such, our model may over estimate the predictive accuracy of digital footprints over political views and our findings/models may not generalize to predicting the ideologies of less politically engaged users whose interests (and hence subreddit interactions) are less tied to their political views. However, these selection effects may actually deflate the observed associations between predictors and response, causing our models to understate the predictive power of digital footprints. The sample selection effects constrain the range of the distributions for our predictors (people who interact with the r/PoliticalCompassMemes subreddit likely have a narrower range of interest and hence subreddit use and language variation than the broader Reddit population) and accordingly reduce the variance in our predictors that can correlate with our response variable.

We note that this sample issue is common in all studies focused on predicting private traits through digital records, many of which feature selection bias by requiring users to voluntarily disclose relevant traits[3, 13–16, 22, 24].

Another very minor potential issue comes from how we collected users' ideology as signalled by their flairs at time t_1 but scraped their most recent comments at time t_2 where $t_2 > t_1$. If a user's ideology has changed between t_1 and t_2 then their digital footprints may not reflect the digital behaviour associated with their flaired ideology, undermining any predictive models. We do not consider this to be a substantive risk given the relatively short interval between the creation of comments used to map a user to a flair and the collection of digital footprints as the subreddit has only been popular for a few years.

To reiterate, censoring of data and non representativeness is common to almost all of the work which uses digital data to predict personal characteristics and whilst our study is flawed by these common disadvantages our data also features some advantages compared to earlier studies.

3.5 Advantages

The previous section elaborates on the faults in our data that are, by and large, endemic to all research in this field. However, our data also promises several advantages over

previous works. We will describe the advantages of our data with reference to the data used by Kosinski et al.[3] since this work is the methodological basis of our paper.

As noted, we want our data to proxy a broader range of digital footprints so that we can understand just how salient digital privacy concerns are. Facebook Likes and Reddit interaction data are both reasonable proxies as they illustrate interest in certain topics and thus perform a function similar to web search logs, purchase records, and other digital footprints. However, we propose that Reddit data better emulates these footprints for several reasons.

Firstly, due to the anonymity of Reddit relative to Facebook, users may be more willing to implicitly disclose private interests via digital behaviour (users may comment in certain pornographic subreddits but few people would willingly disclose this information via their Facebook Likes). These kinds of interests may be connected to ideology. The fact that Reddit data captures this kind of information makes it a better proxy: consider how people typically use their search engine under the presumption that their search queries are private, or how people's purchase records are not broadcast to their friends and families. Essentially, Reddit data expresses the same range of interests as other digital footprints whereas Facebook data only expresses a restricted range of interests. As such, Reddit data should lead to a better indication of the predictive power of digital footprints analogous to search logs, etc. on private traits.

Further, the MyPersonality data set which was used in Kosinski's seminal work[3] and many other studies in the field relied on users reporting their political views on their Facebook profile (linked to their real life account). As such, their data may suffer from a social desirability, 'shy tory' bias where people with ideologies that may be viewed unfavourably by their peers decline to reveal them. The anonymity of Reddit means that r/PoliticalCompassMemes users do not have to worry about their ideologies being linked to their true identity and can freely reveal them.

Our data also features a more sophisticated response variable. Our ideology variable has several categories corresponding to a two dimensional, conception of ideology. In contrast, the majority of work cited in Section 2.1 uses a binary response variable to represent ideology (i.e. liberal/conservative, left/right, Democrat/Republican). This is what allows us to make novel insights regarding the robustness of predictive accuracy when we include ideological centrists and conduciveness of digital footprints to the estimation of different facets of ideology.

Finally, our data set has 91,000 observations[2] and is thus substantially larger than Kosinski et al.'s which featured 9,752 observations with political labels.

[2]In practice, some observations are removed from the training/testing sets for various reasons. The full amount of observations used in training, validating and testing each models is reported alongside metrics of model performance in Section 5.1.

Chapter 4

Methodology

We ran multiple learning models to develop a comprehensive sense of the predictive power of digital footprints over political ideology.

Firstly, we mapped the original flairs to left, center, right and removed centrists to perform binary classification on users' economic ideology. We also mapped flairs to authoritarian, center, libertarian, and removed centrists to perform binary classification on users' social ideology. We refer to these classification problems as the economic problem and social problem respectively. Exact details on how we map from the nine classes to economic classes and social classes can be seen in Appendix B.

The removal of centrist users is justified on the basis that they essentially represent measurement noise. For instance, when we plot the data in singular value decomposition (SVD) component space, we typically see separations between right and left and (to a lesser extent) authoritarian and libertarian users. However, centrist users are randomly splattered throughout the SVD component space indicating that centrist users are not 'true' centrists but have effectively not disclosed their true ideology (recall that a centrist flair is the result of a conscious decision to identify as a centrist rather than as the ideological class suggested by the test). We illustrate this in Figure 4.1 and provide further demonstration in Appendix E.

Further, the existence of 'true' centrists is debated in the political science literature; consider Duverger's proclamation: "in politics, the centre does not exist"[29]. More concretely, recent political science literature has tested competing hypotheses that attempt to explain why centrist self-identification arises. There is evidence that "placing oneself on the center does not indicate that the individual is 'ideologically centrist'"[30] but rather that centrist self-identification predominantly arises from individuals having insufficient political knowledge to place themselves on a left right scale (*uninterested*

hypothesis) and individual's self-identification reflecting decisions to vote for centrist parties or candidates (rather than genuine ideological centrism - the *party-component hypothesis*).

The *uninterested hypothesis* clearly does not apply to users in our sample who willingly identify themselves as centrists and have an interest in politics but the *party-component hypothesis* (and the empirical evidence supporting it) provides a compelling case to disregard those who have labeled themselves as centrists in our initial models.

Finally, removing centrists allows for easier comparisons with the work of Kosinski et al.[3] in which the US party affiliation of Facebook users (Democrat or Republican) is predicted, excluding all users that were not labelled as either.

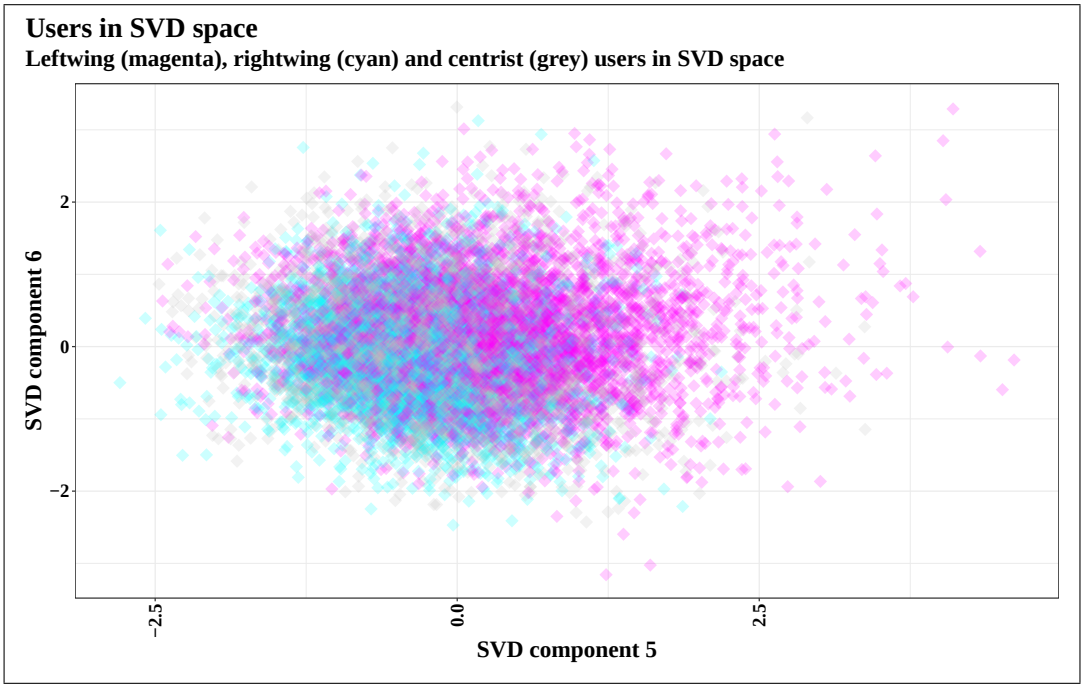

FIGURE 4.1: A scatter plot of users in a subset of SVD space. Colors represent the economic component of users' flairs

After estimating the binary classification models discussed above, we estimate these models again (both for the economic and social problem), this time including centrists. This multi-class classification problem should demonstrate whether or not the results of the binary classifications are robust to the inclusion of so called 'centrists'. Finally, we estimate models for the full nine-class classification problem in which we attempt to classify users to a class from the full set of nine classes.

For each classification problem we run models where we use as features A) the user-interaction matrix, B) features extracted from the textual content of comments and C) the union of these sets of features.

Our data is split into training (64% of total data), validation (16%) and testing (20%) sets. In all models, hyperparameters are chosen according to accuracy on the validation set. We then retrain the model on the training and validation data using optimal hyperparameters. We report the accuracy and where applicable, (weighted) ROC-AUC (receiver operating characteristic, area under curve) of all models' predictions on our testing set and present these results as indicators of the models' predictive power over ideology. Model accuracy and ROC-AUC are expounded in Section 4.4.

Supervised classification tends to become harder as the number of classes increases so we expect accuracy to decrease as classes increase. As such we assess our models' performance against a baseline. We assess the accuracy of our models relative to the baseline accuracy of a ZeroR classifier, which assigns all observations to the plurality class of data in the training set. For example, in the nine-class classification problem, the ZeroR classifier assigns all observations to 'libleft' ($\sim$ 20% of all observations in sample). If digital footprints can predict ideology then our models will exhibit accuracy greater than the accuracy of the ZeroR classifier on the testing set. The more accurate our models are over the ZeroR baseline, the stronger the evidence that digital footprints can predict ideology. ROC-AUC has a natural baseline of 0.5. This is the minimum threshold that a model must surpass to be considered at all useful.

Full specifications of all models for which results are reported (in terms of Scikit-Learn implementation) can be found in C.

4.1 Methodology: user-interaction matrix

When using the user-interaction matrix as our set of features we first remove popular, explicitly political subreddits from the user-interaction matrix prior to running the models. We also trim down our user-interaction matrix to ensure that every user has made at least 50 comments and every subreddit has been commented in at least 50 times as suggested in Kosinski[31]. This also ensures our user-interaction data better proxies digital footprints like search logs which are presumably expansive records. In many cases, we 'binarize' our data, i.e. for each cell (i, j) of the user-interaction matrix, if (i, j) is greater than 0 it is mapped to 1 in the binarized user-interaction matrix and 0 otherwise. The choice to binarize is made on the basis of validation set performance; for most models binarization drastically increased performance.

In developing these models we then follow the same broad approach as Kosinski et al.[3]:

1. Reduce the dimensions of our predictors via a SVD of our data matrix. Many supervised learning methods we wish to apply require less features than observations and a smaller feature set is generally conducive to better performing models.

2. Train a model on the SVD components to classify the ideology of a user.

In step 2 we use the following learning models for multi-class classification: one-versus-rest (OVR) logistic regression (with and without an ℓ_1 penalty), multinomial logistic regression (with and without an ℓ_1 penalty), random forests, OVR random forests and AdaBoost (in the binary case we exclude OVR schemes from all models and multinomial logistic regression reduces to standard logistic regression). A brief account of each model and the SVD is given in 4.5. We used the Scikit-Learn Python package[32] to implement these models.

4.2 Methodology: text

We utilize two approaches to extract features from the textual content of users' comments. Firstly, we create features corresponding to the term frequency-inverse document frequency of different terms in the concatenation of users' comments (tf-idf), where documents are the sets of different users' comments. Secondly, we use a Word2Vec model[33, 34] that has been pre-trained on a Google News data set ('word2vec-google-news-300'[1]) to convert the set of comments for each user into the simple average of the vector representations of their constituent words.

We then classify users' ideologies using a linear support vector classifier (SVC), applying a OVR scheme for multi-class problems. We restrict our choice of supervised learning algorithm to the linear SVC given that the strong performance of support vector machines in document classification is well established[35] (we do not experiment with non-linear kernel functions owing to computational constraints).

We run the SVC using A) tf-idf features, B) average Word2Vec features and C) the union of these two sets of features. Some models use a subset of SVD components from the SVD of original features instead of the original features themselves (this decision is made on the basis of validation set performance). Again, full model specifications can be see in Appendix C.

We outline tf-idf, Word2Vec as well as the linear SVC model in Section 4.5.

[1] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?resourcekey=0-wjGZdNAUop6WykTtMip30g

4.3 Methodology: combined

These models combine the (often binarized) user-interaction matrix and the tf-idf vectors for each user. We scale this data to have the same range since tf-idf terms and user-interaction records are not on the same scale and then take an SVD decomposition. We then feed in the SVD components of this data to an OVR logistic regression model with an ℓ_1 penalty and an OVR linear SVC (or just a simple logistic regression and linear SVC when dealing with binary classification problems).

We do not use the full range of models experimented with earlier because OVR logistic regression with an ℓ_1 penalty typically performed very well for models taking the user-interaction matrix as features. Further, the union of tf-idf and average Word2Vec features leads to (at best) marginally better results than models that just use tf-idf features.

Full model specifications found in C.

4.4 Model assessment

4.4.1 Accuracy

The accuracy of a model on a test set indexed by $i = 1, \ldots, n$ is given by:

$$\text{accuracy} = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i)$$

Where $I(.)$ is an indicator function that returns 1 if the argument inside it is true and 0 otherwise, y_i refers to the true ideology of user i, $\hat{y}_i$ refers to the predicted ideology of user i given by our model. Consequently, accuracy represents the proportion of users in a test set whose ideology our model correctly classifies.

4.4.2 ROC-AUC

To calculate the OVR ROC curves for a multi-class classification model's predictions on our test set we need, for each class k ($k = 0, \ldots, K-1$), estimates of the probability that each observation in the training set, i, is in class k: $\hat{P}(y_i = k)$. Then, for a set of thresholds, $\tau \in [0, 1]$, we assign observation i to k if $\hat{P}(y_i = k) > \tau$ and to 'not

k' otherwise. We then compute the sensitivity (true positive rate) and 1−specificity (specificity is the true negative rate):

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\sum_{y_i=k} I(\hat{P}(y_i = k) > \tau)}{\sum_{i=1}^{n} I(y_i = k)}$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} = \frac{\sum_{y_i \neq k} I(\hat{P}(y_i \leq \tau))}{\sum_{i=1}^{n} I(y_i \neq k)}$$

We plot the values for sensitivity and 1−specificity in ⟨1-specificity, sensitivity⟩ space for each τ. The area under this curve is the AUC for class k; roc auc_k. We compute the AUC for each class and weight it by the total proportion of testing observations in class k; p_k. We sum these K weighted, class specific ROC-AUC values to gain our overall testing ROC-AUC estimate for our model[36]:

$$\text{roc auc} = \sum_{k=1}^{K} \text{roc auc}_k \cdot p_k$$

The OVR AUC for class k is equivalent to the probability that the classifier will assign a randomly selected observation from class k a higher probability of membership in class k than a randomly selected observation from some class other than k[36]. Thus, it is essentially a measure of how well our classifier can distinguish between classes, weighted by class prevalence. This method of averaging OVR AUC scores is referred to as the 'weighted' method in SciKit-Learn[32].

For binary classification problems we simply compute the probability that each observation belongs to one of the two classes (left/right or auth/lib) and compute the sensitivity and $1 -$ specificity associated with varying thresholds to construct the ROC curve and therefore determine ROC-AUC.

ROC curves can only be constructed for models that can estimate the probability that an observation belongs to a class. As such, we omit the ROC-AUC metric from the evaluation of some models. Exactly how probability estimates are obtained from each model can be seen in the Scikit-Learn documentation[32].

4.5 Overview of statistical learning approaches

4.5.1 Singular value decomposition

Singular value decomposition (SVD) is a dimension reduction technique that allows us to replace our data matrix, $X_{(n\times p)}$, with a lower dimensional representation: a matrix $\bar{X}_{(n\times q)}$ where $q < p$. The SVD computation finds the p dimensional vector, $\vec{v_1}$, in feature space that minimizes projection error of data onto the vector. This is the first SVD component. We can then represent each data point by its projection onto this vector. The distribution of the projected values of data points has the largest variance of any possible projection. We then determine subsequent SVD components by finding the p dimensional vector that maximises the variance of projected data points subject to the constraint that the vector is orthogonal to existing SVD components. We do this until we have q vectors.

Mathematically, we are decomposing the data matrix X into the product of three matrices:

$$X_{(n\times p)} = U_{(n\times r)}\Sigma_{(r\times r)}V^{\intercal}_{(r\times p)}$$

Here, U is an $n \times r$ matrix where the n rows correspond to users and the r columns correspond to 'themes' in the subreddit interaction data; i.e. latent factors. Any particular cell $u_{(i,j)}$ intuitively represents how much user i interacts with theme j. Σ is a $(r \times r)$ diagonal matrix where $\sigma_{(i,i)}$ represents the variance of theme i. V is a $(p \times r)$ matrix that, when transposed, maps interaction with subreddits to themes, i.e. $V^{\intercal}_{(i,j)}$ represents how much each interaction with subreddit j adds or detracts from the the score on theme i. $\vec{v_q}$ is the q_{th} row of the matrix $V^{\intercal}$.

We choose to use the first 500 SVD components as features in our model. In conjunction with trimming, this reduces the ammount of features in the user-interaction matrix from over 190,000 to 500. This accounts for roughly 86% of the variance of the original data.

As noted by Kosinski, principal component analysis is a less appropriate technique for the kind of problem we are dealing with as it requires a larger amount of computational resources. It also requires that the data be centered which does not preserve sparsity and efficient computation[31].

4.5.2 OVR logistic regression

We first describe a simple logistic regression (as used in our binary classification problems) and then explain how it can be extended to multi-class classification problems through a OVR scheme.

A logistic regression models assumes that, when individual i makes the binary decision $y_i \in \{0, 1\}$, the probability they choose $y_i = 1$ is given by:

$$P(y_i = 1|x_i;\, \beta) = \frac{e^{x_i^\intercal \beta}}{1 + e^{x_i^\intercal \beta}}$$

Where x_i is a vector of predictors (in our case, 500 SVD components) with $x_{i,0} = 1$ to allow for a constant term. $\beta = (\beta_0, \beta_1, \ldots, \beta_{500})^\intercal$ is a vector of coefficients to be estimated. Estimated model coefficients, β^* are obtained via maximum likelihood estimation (MLE), i.e. through solving the following maximization problem[37] (where $i = 1, \ldots, n$ now indexes observations in our training set):

$$\beta^* = \underset{\beta}{\operatorname{argmax}} \left\{ \frac{1}{n} \sum_{i=1}^{n} w_i \{ y_i \ln P(y_i = 1|x_i;\, \beta) + (1 - y_i) \ln(1 - P(y_i = 1|x_i;\, \beta)) \} \right\}$$

w_i is the weight for each observation. We try both a uniform weighting and balanced weighting: $w_i = \frac{n}{n_i}$ where n_i refers to the number of observations in the same class as observation i.

With our MLE estimate of β, $\hat{\beta}$, we can can estimate the probability that some person j, who does not feature in our training set, chooses $y_j = 1$ as:

$$\hat{P}(y_j = 1|x_i;\, \hat{\beta}) = \frac{e^{x_j^\intercal \hat{\beta}}}{1 + e^{x_j^\intercal \hat{\beta}}}$$

This works for binary dependent variables but we are often attempting to classify ideologies where our dependent variable, y_i, takes on one of the following forms:

- $y_i \in$ {libleft, libright, libcenter, centrist, left, right, authright, authleft}
- $y_i \in$ {left, center, right}
- $y_i \in$ {authoritarian, center, libertarian}

We can extend the binary model using a technique called 'one versus rest' (OVR) logistic regression which involves fitting a model for each class to predict the probability that someone is in that class versus not in that class.

Consider the economic-ideology classification problem where $y_i \in \{\text{left, center, right}\}$. The OVR logistic regression approach to modelling this problem involves fitting three different logistic regression models.

Model 1 - predict probability that individual is left versus not left:

$$\hat{P}_1(y_i = \text{left}|x_i;\, \hat{\beta}) = \frac{e^{x_j^\intercal \hat{\beta}}}{1 + e^{x_j^\intercal \hat{\beta}}}$$

Model 2 - predict probability that individual is center versus not center:

$$\hat{P}_2(y_i = \text{center}|x_i;\, \hat{\gamma}) = \frac{e^{x_j^\intercal \hat{\gamma}}}{1 + e^{x_j^\intercal \hat{\gamma}}}$$

Model 3 - predict probability that individual is right versus not right:

$$\hat{P}_3(y_i = \text{right}|x_i;\, \hat{\delta}) = \frac{e^{x_j^\intercal \hat{\delta}}}{1 + e^{x_j^\intercal \hat{\delta}}}$$

We assign individual j to the class for which the relevant OVR probability expression is largest. For example, we assign j to left iff:

$$\frac{e^{x_j^\intercal \hat{\beta}}}{1 + e^{x_j^\intercal \hat{\beta}}} > \frac{e^{x_j^\intercal \hat{\gamma}}}{1 + e^{x_j^\intercal \hat{\gamma}}}, \text{ and } \frac{e^{x_j^\intercal \hat{\beta}}}{1 + e^{x_j^\intercal \hat{\beta}}} > \frac{e^{x_j^\intercal \hat{\delta}}}{1 + e^{x_j^\intercal \hat{\delta}}}$$

4.5.3 OVR logistic regression with ℓ_1 penalty

To improve upon the basic logistic regression model, we can augment the log likelihood by subtracting the ℓ_1 norm (sum of absolute value) of the coefficient vector from the log likelihood. Thus that we obtain coefficient estimates through solving the following maximization problem:

$$\beta^* = \underset{\beta}{\text{argmax}} \left\{ \frac{1}{n} \sum_{i=1}^{n} w_i \{ y_i \ln P(y_i = 1|x_i;\, \beta) + (1 - y_i) \ln(1 - P(y_i = 1|x_i;\, \beta)) \} - \lambda ||\beta_{/0}||_1 \right\}$$

Where $||\beta_{/0}||_1$ refers to the ℓ_1 norm of the $(p \times 1)$ vector $\beta_{/0}$, which is the vector of coefficients associated with all regressors excluding the intercept term. i.e. $||\beta_{/0}||_1 = \sum_{j=1}^{p} |\beta_j|$.

When coefficients deviate from zero in magnitude, $\lambda||\beta_{/0}||_1$ increases, thus penalising the augmented log-likelihood. This has the effect of shrinking unimportant coefficients to zero and may improve predictive accuracy if the decrease in model variance more than offsets the increase in model bias[38]. The lasso penalty can be viewed as the imposition of a constraint on the size of parameters onto the maximisation problem discussed in Section 4.5.2. The parameter λ is a tuning parameter that controls the strength of the regularization; a higher λ implies stronger regularization i.e. a smaller budget for the maximum ℓ_1 norm of the coefficient vector.

This process provides a different set of coefficient estimates (with many possibly being equal to zero, effectively dropping the irrelevant features from the model) but the process of prediction and implementation of the OVR scheme is the same as in Section 4.5.2.

4.5.4 Multinomial logistic regression

Multinomial logistic regression is an extension of logistic regression that can naturally perform multi-class classification. When there are K choices $(k = 0, \ldots, K-1)$, the probability that individual i chooses choice k is given by:

$$P(y_i = k|x_i,\ \beta_k) = \begin{cases} \dfrac{1}{1 + \sum_{k\prime=1}^{K} e^{x_i^\intercal \beta_{k\prime}}} & k = 0, \\ \dfrac{e^{x_i^\intercal \beta_k}}{1 + \sum_{k\prime=1}^{K} e^{x_i^\intercal \beta_{k\prime}}} & k \neq 0 \end{cases}$$

Thus, prediction requires the estimation of a separate coefficient vector β_k for each class. By convention, β_0 is set to $\vec{0}$ since we cannot uniquely identify any particular β_k, we can only identify $(\beta_k - \beta_0)$[37].

Coefficient vector estimates, $\hat{\beta}_1, \ldots, \hat{\beta}_{K-1}$, are found by solving the following maximization problem:

$$\hat{\beta}_1, \ldots, \hat{\beta}_{K-1} = \underset{\beta_1, \ldots, \beta_{K-1}}{\operatorname{argmax}} \left\{ \frac{1}{n} \sum_{i=1}^{n} w_i \sum_{k=0}^{K} I(y_i = k) \ln P(y_i = k|\beta_k) \right\}$$

This allows us to compute the predicted probability that a new observation, j, belongs to any particular class like so:

$$\hat{P}(y_i = k|x_i;\ \hat{\beta}_k) = \begin{cases} \dfrac{1}{1 + \sum_{k\prime=1}^{K} e^{x_i^\intercal \hat{\beta}_{k\prime}}} & k = 0, \\ \dfrac{e^{x_i^\intercal \hat{\beta}_k}}{1 + \sum_{k\prime=1}^{K} e^{x_i^\intercal \hat{\beta}_{k\prime}}} & k \neq 0 \end{cases}$$

We assign j to $\underset{k}{\text{argmax}}\, \hat{P}(y_j = k|\hat{\beta}_k)$ to obtain a categorical prediction.

4.5.5 Multinomial logistic regression with ℓ_1 penalty

The ℓ_1 penalty described in Section 4.5.3 can also be fruitfully applied to the multinomial logistic regression. Thus, we are choosing $\hat{\beta}_1, \ldots, \hat{\beta}_{K-1}$ to maximise[37]:

$$\hat{\beta}_1, \ldots, \hat{\beta}_{K-1} = \underset{\beta_1,\ldots,\beta_{K-1}}{\text{argmax}} \left\{ \frac{1}{n} \sum_{i=1}^{n} w_i \sum_{k=0}^{K} I(y_i = k) \ln P(y_i = k|\beta_k) - \lambda \sum_{k=1}^{K} ||\beta_{k,/0}||_1 \right\}$$

Where $\beta_{k,/0}$ refers to the coefficient vector β_k without the intercept coefficient.

This has the same effect as described in Section 4.5.3[37]. The rest of the classification procedure after model estimation is the same as described in Section 4.5.4.

4.5.6 Random forests and OVR random forests

A random forest classifier is a a collection of uncorrelated classification trees. We use each of the constituent trees to predict a users ideology and assign users to the ideology which received a plurality of votes from the set of tree models. Each constituent tree may only use a random subsample of the full range of predictors at a given split to ensure that trees remain uncorrelated[38]. Using an ensemble of tree predictors provides a model with lower variance than an individual classification tree[38]. Each tree uses a bootstrapped sample of size n as its training set.

A classification tree recursively partitions the p dimensional feature space into non-overlapping subsets. Binary partitions are made along a single variable to minimize some quantity measuring the impurity (heterogeneous class membership) of the resulting partitions. The splitting process stops when certain criterion are reached. Predictions are made by assigning a new observation to to the plurality class for the partition in feature space in which the observation resides.

The trees in our random forest model make partitions to minimise Gini Impurity[38]. The Gini Impurity for a given partition, m is given by:

$$G_m = \sum_{k=1}^{K} p_{mk}(1 - p_{mk})$$

Where p_{mk} refers to the proportion of observations in partition m belonging to class k. We experiment with uniform sample weights, w_i and balanced weights where w_i are assigned to be inversely proportional to class frequency within the relevant bootstrap sample, i.e. $w_i = \frac{n_B}{n_{B,i}}$ where n_B is the number of samples in the relevant bootstrap sample and $n_{B,i}$ is the number of samples of the same class as observation i in n_B. Consequently, $p_{mk} = \dfrac{\sum_{x_i \in m} w_i I(y_i = k)}{\sum_{x_i \in m} w_i}$.

Consider a candidate split of our feature space X into: $P_1 = \{X : x_p < a\}$ and $P_2 = \{X : x_p \geq a\}$. The Gini Impurity of the resulting partitions is given by:

$$n_1 \cdot G_1 + n_2 \cdot G_2$$
$$= n_1 \cdot \sum_{k=1}^{K} p_{1k}(1 - p_{1k}) + n_2 \cdot \sum_{k=1}^{K} p_{2k}(1 - p_{2k})$$

Where n_1 refers to the number of observations in P_1; $|\{X : x_p < a\}|$, and n_2 refer to the number of observations in P_2; $|\{X : x_p \geq a\}|$. x_p is chosen from the set of predictors and a is chosen from the set of values in the range of x_p to minimise Gini Impurity.

We specify a minimum bucket and minimum samples stopping criteria. The minimum bucket specification requires that a minimum number of samples be included in each partition. The minimum sample criterion requires that a partition with less samples than the minimum not be split again.

Random forests are capable of multi-class classification but an OVR estimation scheme as detailed in Section 4.5.2 can also be applied to random forest models and can at time outperform standard random forests[39]. Continuing with the economic-ideology classification problem as an example, $y_i \in \{\text{left, center, right}\}$, we would implement an OVR random forest scheme by training 3 models: a random forest that predicts being left versus not left, center versus not center and right versus not right.

4.5.7 AdaBoost

AdaBoost is a boosting algorithm that produces a classifier from a set of weak classifiers (in our case, decision trees). A sequence of stumps, $T^{(1)}, \ldots, T^{(M)}$ (decision trees of depth 1) are created where the errors of m_{th} tree impact the structure of the $m + 1_{\text{th}}$ tree. Initially, all samples are weighted equally. The first model is the decision stump that best partitions the data according to the Gini index (see Section 4.5.6). The stump's input to the final classification decision is determined by the accuracy with which the stump partitions the training set (impurity metrics account for each samples weight). Finally, the sample weights of observations that were incorrectly classified are increased whilst the weights of those correctly classified are decreased. The magnitude of increases and decreases to sample weights depends on the weight of the most recently trained constituent stump. The next stump is defined by the split that yields the strongest weighted accuracy in its partitioning of the data. This process continues until we have M classifiers working in unison. A technical description of the SAMME multi-class AdaBoost algorithm[40] is provided below:

Training:

1. Initialise observation weights, w_i: $\forall i \in [1,\, n] : w_i := \frac{1}{n}$
2. For $m = 1, \ldots, M$:
   (a) Fit tree stump $T^{(m)}(x_i)$ to training data using Gini index on observations with observations weighted by w_i ((weighted Gini)).
   (b) Calculate the error of $T^{(m)}(x_i)$: $\epsilon^{(m)} = \dfrac{\sum_{i=1}^{n} w_i I(y_i \neq T^{(m)}(x_i))}{\sum_{i=1}^{n} w_i}$.
   (c) Calculate the weight of $T^{(m)}(x_i)$'s classification in the ensemble classifier where $\eta \leq 1$ is the specified learning rate: $\alpha^{(m)} = \eta \cdot \log \dfrac{1 - \epsilon^{(m)}}{\epsilon^{(m)}} + \log(K - 1)$.
   (d) Update the observation weights: $\forall i \in [1,\, n] : w_i := w_i \cdot e^{\alpha^{(m)} I(c_i \neq T^{(m)}(x_i))}$.
   (e) Re-normalise observation weights: $\forall i \in [1,\, n] : w_i := \dfrac{w_i}{\sum_{j=1}^{n} w_j}$.

Prediction:

To estimate the class, $C(.)$, of a new observation, j, we take the weighted vote of our M tree stumps (votes are weighted by classifier weight $\alpha^{(m)}$) and assign observation j to the class k which received a plurality of votes:

$$C(x_j) = \underset{k}{\text{argmax}} \sum_{m=1}^{M} \alpha^{(m)} I(T^{(m)}(x_j) = k)$$

Rojas provides lighter explanation of the AdaBoost algorithm for binary classification tasks[41].

4.5.8 OVR linear support vector classifier

We describe a linear support vector classifier (SVC) for a binary classification problem. This is extended to multi-class classification via a OVR scheme that determines an observation's class on the basis of distance between several, class specific separating hyperplanes.

The SVC estimates a linear decision boundary in feature space by finding the hyperplane that maximizes the margin (euclidean distance between the hyperplane and the observations closest to the hyperplane from either class) subject to a certain tolerance for training observations that violate the margin (i.e. that are on the 'wrong' side of the margin).

The coefficients for this hyper plane, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_{/0}^{\intercal})^{\intercal}$ are found by solving the following problem[38]:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} = \frac{1}{2}\beta_{/0}^{\intercal}\beta_{/0} + C\sum_{i=1}^{n} \max(0,\, y_i(\beta_{/0}^{\intercal}x_i + \beta_0))$$

Here x_i does not include a constant first element equal to one. C is a constant that is *inversely* proportional to the strength of regularization[2]. Higher values of C will correspond to a stricter margin with less tolerance for violation.

We can predict the class of an observation, j depending on which side of the estimated separating hyper plane it falls on, i.e depending on the sign of $x_j^{\intercal}\hat{\beta}$.

4.5.9 Term frequency - inverse document frequency

The first step in this process of feature extraction is to 'clean' the text. We remove all html links and punctuation, non ascii symbols (i.e. emojis) and single letter terms, convert all terms to lower case, and finally stem all terms.

[2] In some formulations of the linear SVC maximization problem C refers to the regularization parameter itself but we follow the formulation given in the Scikit-Learn documentation.

Then, for each word, w we compute the 'term frequency' for user i:

$$TF_{w,i} = \frac{n_{w,i}}{\sum_{k=1}^{K} n_{k,i}}$$

Where $n_{w,i}$ refers to how many times word w appears in user i's concatenated comments and $k = 1, \ldots, K$ indexes all words appearing in all comments for each user.

When then compute the 'inverse document frequency' of each word[3]:

$$IDF_w = \ln\left(\frac{D+1}{D_w+1}\right) + 1$$

Where D refers to the total users whose comments we have recorded (total amount of documents) and D_w is the amount of concatenated sets of comments containing word w.

Consequently, $TF_{w,i} \cdot \ln\left(\frac{D+1}{D_w+1}\right) + 1$ measures how much user i uses word w relative to other users. We construct inverse document frequency for every word in our training vocabulary (set of terms used in comments from users in our training set) and then create tf-idf scores for each user on all of these terms (both in the training and testing set). We then normalize tf-idf vectors for each user by their euclidean norm.

Several hyperparameters impact the set of terms for which we compute tf-idfs, namely the maximum and minimum frequency of words with which we allow in our vocabulary (words that occur in a greater proportion of documents than our maximum frequency or in a smaller proportion of documents than our minimum frequency are disregarded). We also specify the maximum amount of words in the vocab, where the top words are chosen according to their term frequency across the corpus.

4.5.10 Word2Vec

Word2Vec is an unsupervised learning algorithm that takes a corpus of text and 'learns' to represent each word as a vector. We use embedding from a model that has been pre-trained on Google News data. This model contains 300 dimensional vectors for around 100,000,000,000 words and does not require that we convert our text to lower case or remove punctuation.

[3]This specific form is implemented by our models and differs slightly from standard textbook definitions of inverse document frequency.

Using a pre-trained model allows us to avoid training our own model which is time consuming and may be inaccurate due to our relatively small corpus of text.

The workings of the Word2Vec algorithm are somewhat complex so for the save of brevity we refer the reader to [33] and [34].

Chapter 5

Results

In Section 5.1 we outline the performance of the models discussed in Section 4. Section 5.2 contains some interesting visualisations from our data.

5.1 Model results

Note: *in this section, we discuss the performance of the best models for each classification task. Best model refers to the model with the highest testing accuracy. This is not necessarily the model with the highest ROC-AUC. Further, as discussed, not all models features an ROC-AUC value. Specifcally, the linear SVC is incapable of computing probabilities so reports of ROC-AUC are omitted for these models.*

The accuracy and ROC-AUC (where applicable) of every model estimated can be found in tables XX to YY.

In the binary classification of economic ideology (i.e. $y_i \in$ {left, right}) using only the user-interaction matrix as features, our best model's (logistic regression) predictions on the test set achieved an accuracy of 82.4 % (against a ZeroR baseline of 56.1 %) and an ROC-AUC of 91%. This reinforces Kosinski et al's findings. Kosinski et al. were able to predict whether a Facebook user was a Democrat or Republican based off their Facebook Likes[3] (nothing that they do not omit Facebook Likes for explicitly political pages) with an ROC-AUC of 85%. Our model is able to better predict ideology, further validating these findings.

In contrast, our best model in the binary classification of social ideology (i.e. $y_i \in$ {lib, auth}) using only the user-interaction matrix, logistic regression, barely improved

upon the ZeroR baseline accuracy of 79%, achieving an accuracy of 82% on testing data. The model achieved an ROC-AUC of 78.7% on the test set[1].

Textual features performed somewhat worse in the binary classification problems. In predicting economic ideology, the linear SVC applied to the union of tf-idf features and average Word2Vec features (which was the best model in this task) achieved an accuracy of 73.2% relative to the 54.1% ZeroR baseline. For social ideology, the best model, a linear SVC using just tf-idf features, achieved an accuracy of 80% against an 76.6% baseline[2]. Interestingly, this illustrates that the difficulty of predicting social ideology relative to economic ideology is robust to the type of digital record employed.

The models that utilised both the user-interaction matrix and textual footprints (tf-idf features) did not substantially improve upon the performance of the models that solely used the user-interaction matrix as features.

We now report the performance of models in the economic and social ideology classification tasks where we do not omit centrist observations, i.e. we classify $y_i \in \{\text{left, center, right}\}$ and $y_i \in \{\text{lib, center, auth}\}$. Using just the user-interaction matrix, in the economic classification task our best model, OVR logistic regression with an ℓ_1 penalty, achieves an accuracy of 57.6% against a ZeroR baseline of 36.6% and an ROC-AUC of 76.5 %. For the social problem, OVR logistic regression marginally surpassed baseline performance, achieving an accuracy of 56% (baseline: 52.8%) and an ROC-AUC of 67.3%.

The models using textual features again perform worse than those using the user-interaction matrix. The best economic classifier using textual features (linear SVC using a union of tf-idf features and average Word2Vec features), still performs substantially better than baseline benchmarks (accuracy: 49% vs 35.1% ZeroR accuracy). The best social model (linear SVC using just tf-idf features) achieves an accuracy of 53.7% against a 51.4% baseline. This indicates that the classification of economic ideology is robust to the inclusion of centrists (i.e. we can still classify users' ideology at an accuracy substantially above baseline), albeit with significantly diminished accuracy.

However, the stronger performance of the user-interaction matrix relative to textual features is also robust to the inclusion of centrists. This is perhaps surprising as, typically, textual features tend to predict ideology to a high degree of accuracy.

We suspect the primary reason for the comparatively poor performance of textual features is due to the limited variation in language use in our sample. The r/PoliticalCompassMemes subreddit is an online subculture with its own vernacular. For instance,

[1] The high ROC-AUC figure here cannot be trusted as the response variable is highly imbalanced as indicated by the ZeroR classification accuracy.

[2] ZeroR baselines slightly differ from those in the models because the set of users used in developing models with textual features and user-interaction features slightly differs.

many users of all ideologies use the jargon 'based' and 'cringe' to indicate appreciation or dissatisfaction. We scraped comments from subreddits outside of r/PoliticalCompassMemes to try and maximise variation in word use but the use of the vernacular likely persists in other subreddits. Further, though we are unable to rigorously validate this claim, most users are likely young; many users in our sample comment in subreddits indicative of age, i.e. r/teenagers and appear to be located in the United States based on the prevalence of discussion topics relevant to US politics.

This effectively restricts the variation in vocabulary used by r/PoliticalCompassMemes commentors making it harder to link variation in language to ideology. The lack of variation in language is shown in Figure 5.1 and Figure 5.2 (note: the frequency of the n-word and 'cum' is due to a handful of 'troll' users who have made many comments which repeat these words thousands of times).

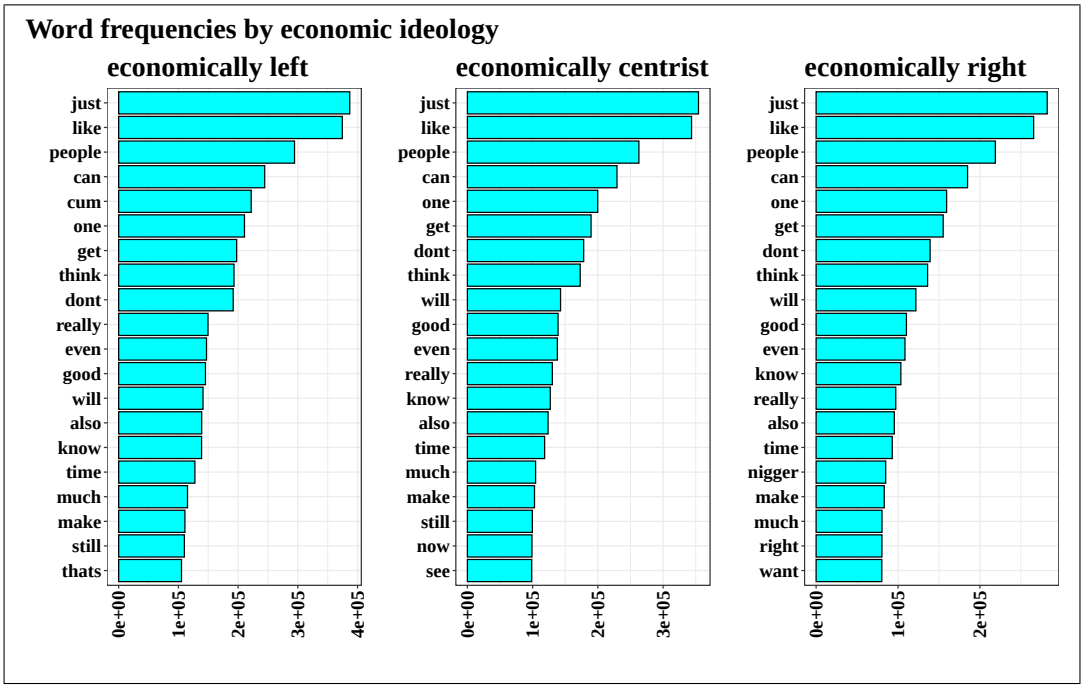

FIGURE 5.1: Word frequencies in comments by economic ideology

Still, variation in language is associated with ideology and textual features remain useful. To provide some insight into which facets of language are most associated with ideology we display the words most strongly associated with economic and social ideologies in Figure 5.3 and Figure 5.4.

Again, the union of the user-interaction matrix and textual features does not lead to models with substantively better performance on unknown data. We suspect that this is due to colinearity between the user-interaction features and textual features. In the same way that r/PoliticalCompassMemes has its own vernacular, interaction with subreddits is likely associated with particular language usage. If a user has made comments

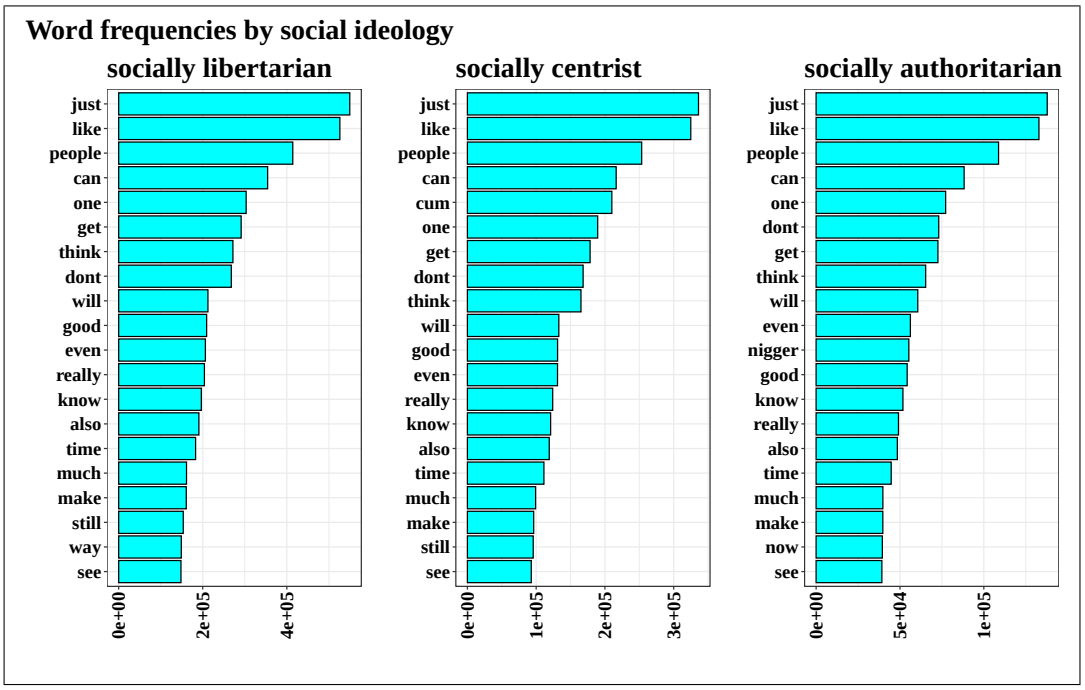

FIGURE 5.2: Word frequencies in comments by social ideology

in the r/Vegetarian subreddit then they have likely used language strongly associated with vegetarianism, i.e. 'vegan', 'meat', 'dairy', increasing the value of textual features associated with use of these words. Thus, engagement with subreddits and language use likely go hand in hand in many cases and there is little information gained from textual features when we have already employed records of subreddit interaction.

The final set of models estimated predict which of the full nine classes each user belongs to, i.e. $y_i \in \{\text{centrist, left, right, libright, libleft, libcenter, authright, authleft, authcenter}\}$. Understandably, these models feature the lowest accuracy due to the abundance of possible classes. Nevertheless, the best model based on user-interaction matrix (multinomial logistic regression) was able to substantially improve upon baseline accuracy achieving an accuracy of 34.8% against the 20.8% ZeroR baseline and an AUC of 74.6%. The best nine-class classification model using solely textual data (linear SVC on union of tf-idf features and average Word2Vec features) perform worse than the best model using solely the interaction data achieving an accuracy of 24.3% against the 19.8% ZeroR baseline. The union of textual features with the user-interaction matrix does not substantially improve performance in the nine-class classification problem.

5.2 Visualisations

Here, we display the proportion of total comments from each ideology in particular subreddits. This is intended to illustrate the variation in subreddit interaction between

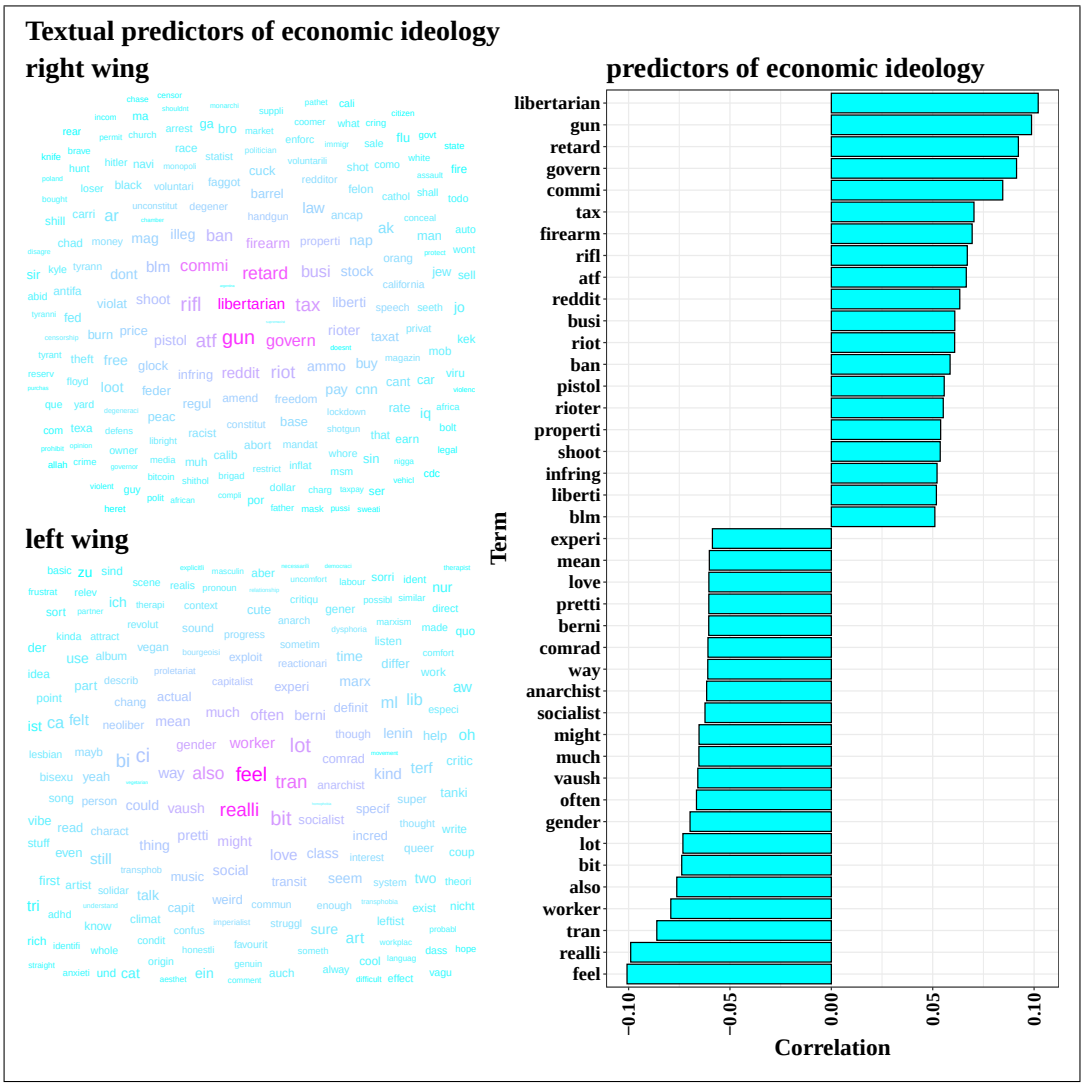

FIGURE 5.3: Economic ideology was recoded from {left, center, right} to {-1, 0, 1}. We then computed the correlation of the tf-idf scores for each term with the quantitative economic ideology measure. The column chart displays the top 20 terms with the strongest positive correlation (terms most associated with users with right wing economic views) and the top 20 terms with the strongest negative correlation (terms associated with users with left wing economic views). Size of terms in wordclouds is proportional to correlation with the relevant ideology

users of different ideologies and illustrate how Reddit data can illuminate psycho-social dynamics.

From Figure 5.5 one could infer that those with more left-leaning economic views are more comfortable discussing any mental illnesses they are facing. This also illustrates how digital footprints related to mental health (i.e. search engine queries relating to a condition) are the type of digital footprint that could provide information on a person's political beliefs despite the fact there is no obvious connection between mental health and political ideology.

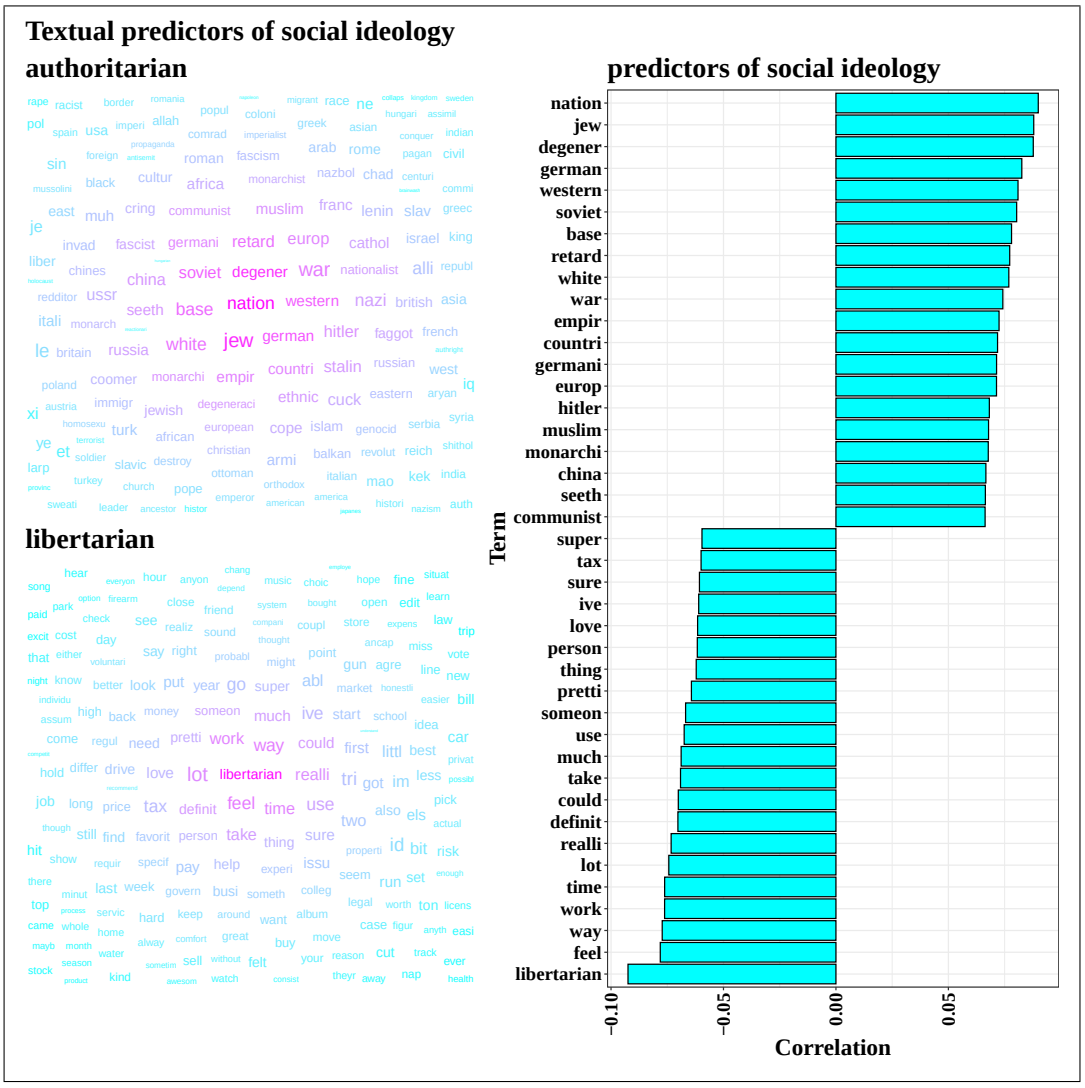

FIGURE 5.4: Social ideology was recoded from {lib, center, auth} to {-1, 0, 1}. We then computed the correlation of the tf-idf scores for each term with the quantitative social ideology measure. The column chart displays the top 20 terms with the strongest positive correlation (terms most associated with users with authoritarian social views) and the top 20 terms with the strongest negative correlation (terms associated with users with libertarian social views). Size of terms in wordclouds is proportional to correlation with the relevant ideology

We can also examine the relationship between ideology and subreddits we might expect to be associated with particular ideologies. Figure 5.6 shows these results for several subreddits that common wisdom dictates would be favoured by users with progressive views. r/lgbt is a subreddit for people with sexual orientations other than heterosexual, r/AgainstHateSubreddits is a forum focused on identifying and denouncing other subreddits perceived to have engaged in prejudiced behaviour and r/TwoXChromosomes is a subreddit for women (studies show women tend to lean further left). Variation in the amount of interaction with these subreddits along ideological lines allows for digital footprints to predict ideology.

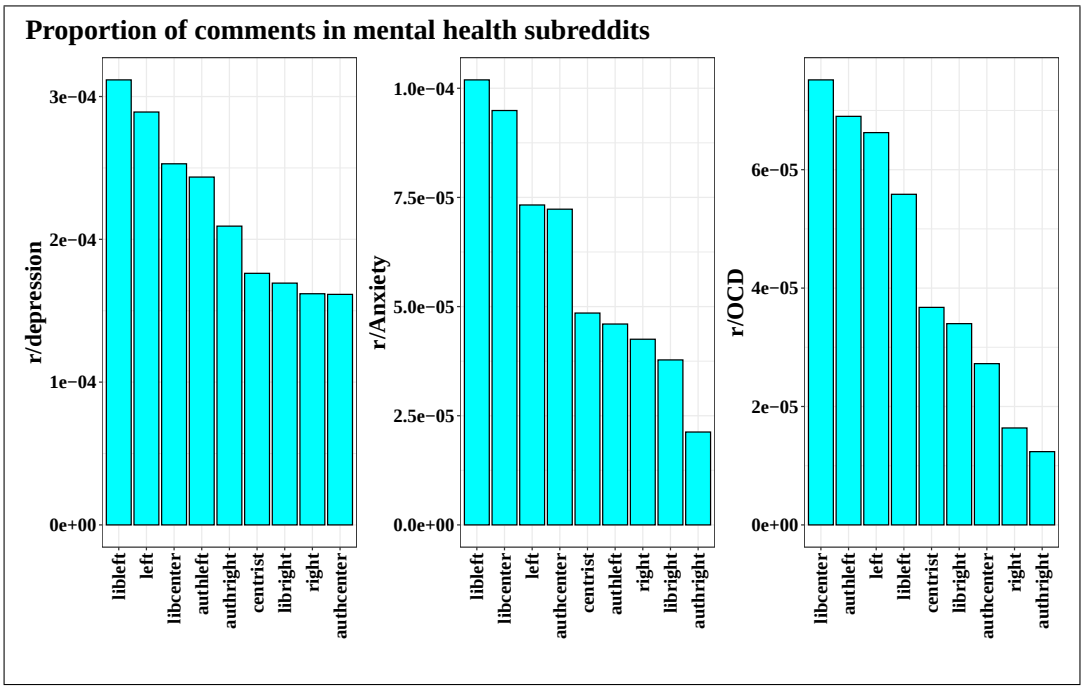

FIGURE 5.5: This figure shows the proportion of total comments from users of each ideology in a range of subreddits related to mental health conditions

Interestingly, the right wing authoritarian users have the largest proportion of comments/posts in r/AgainstHateSubreddits which is contrary to expectation. A brief investigation suggests that this may be because users from r/AgainstHateSubreddits have actually targeted r/PoliticalCompassMemes due to the behaviour of the more conservative users. Consequently, some conservative users comment in r/AgainstHateSubreddits to engage in argument.

Figure 5.7 illustrates how libertarian users of any economic persuasion (but especially those with right wing economic views) interact the most with r/Bitcoin. This makes sense as Bitcoin is a decentralised digital currency that can be used anonymously; which would naturally appeal to those who are skeptical of state power and value political freedom highly.

We also see that economically right wing users interact most with the popular r/wallstreetbets subreddit which is focused on risky trading strategies in financial markets. Economically right wing users also interact more with the r/conspiracy subreddit.

Of even greater interest, and perhaps concern, is the fact that interactions with subreddits that do not pertain to politically charged interests exhibit substantial variation with respect to ideology. For example, as illustrated in Figure 5.8, those with authoritarian views seem to have much less interest in movies or sport; perhaps those with authoritarian views are more committed to their political beliefs and consequently these beliefs form a larger part of their identity leaving them less interested in non-political

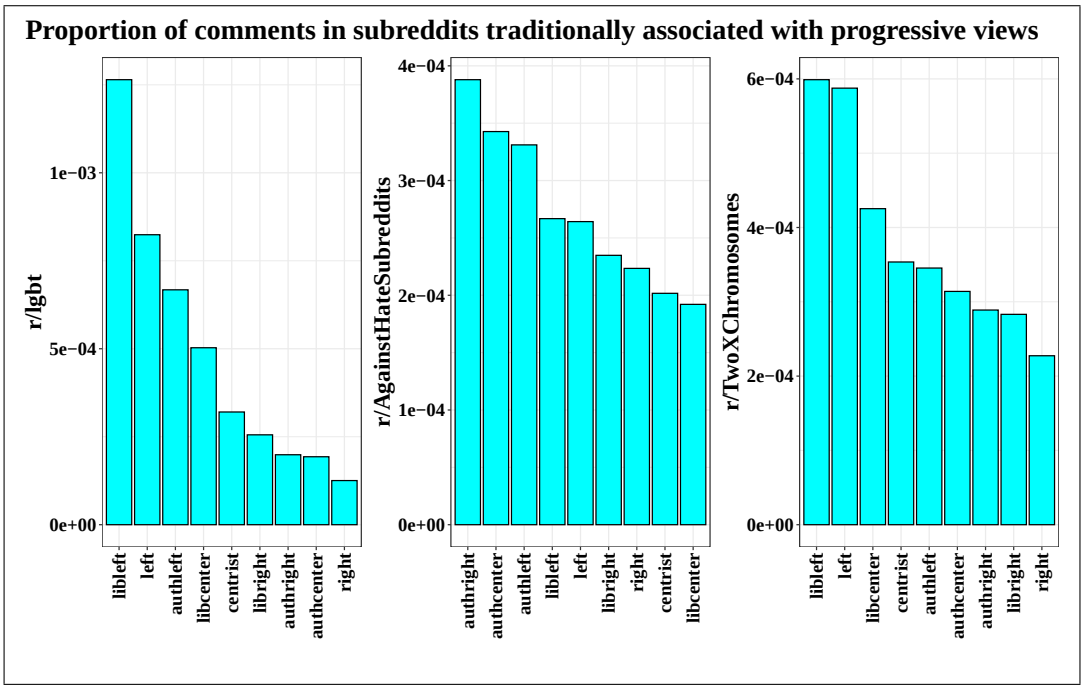

FIGURE 5.6: This figure shows the proportion of total comments from users of each ideology in a range of subreddits that common wisdom suggests are related to progressive political views

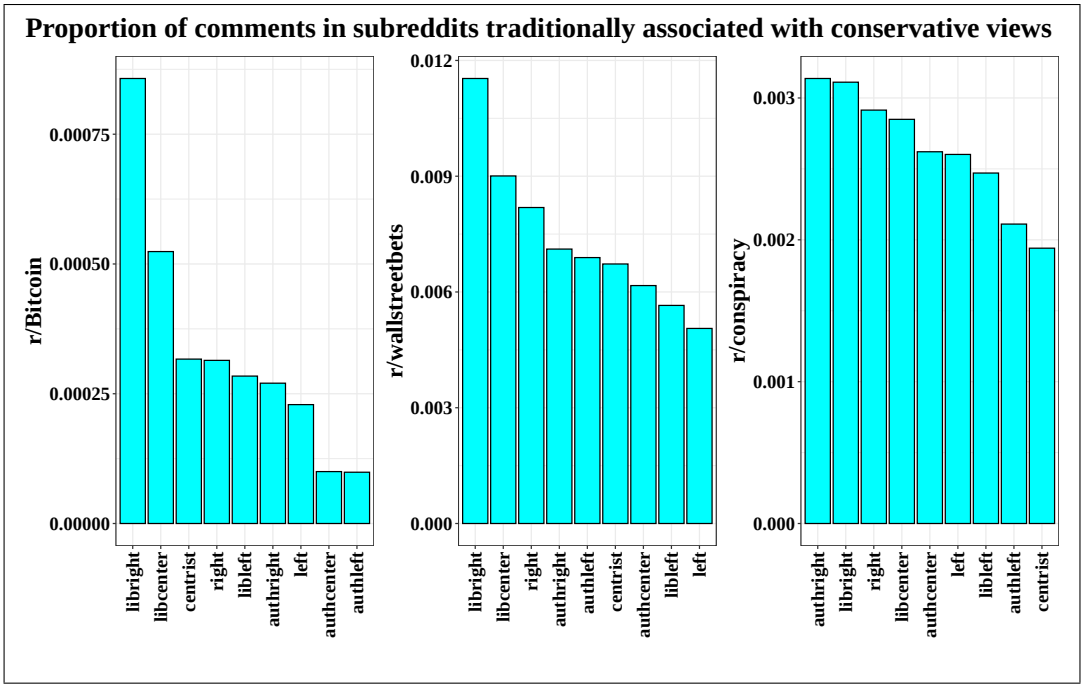

FIGURE 5.7: This figure shows the proportion of total comments from users of each ideology in a range of subreddits that common wisdom suggests are related to conservative political views

activities. The upshot here is that digital footprints indicating something as simple as your interest in movies or sports is informative with respect to ideology.

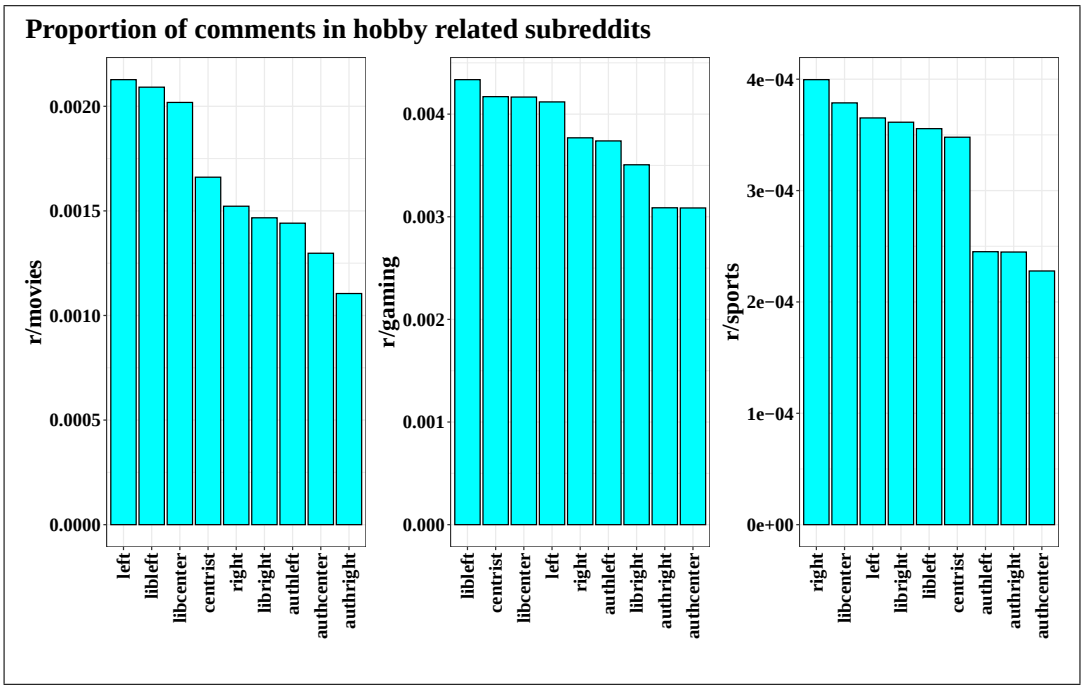

FIGURE 5.8: This figure shows the proportion of total comments from users of each ideology in a range of subreddits that relate to common interests and hobbies

Chapter 6

Discussion

6.1 Policy implications

We have shown that, publicly available, largely non-political digital footprints can be mapped to digital footprints with greater than baseline accuracy. This broader claim provides validation of existing results but we have also made several novel contributions:

1. Showing that digital footprints are able to predict a multi-class ideological label that includes centrists rather than just a simple left/right dichotomy, although this is a trickier, noisier problem.
2. Showing that digital footprints are better predictors of economic ideology than social ideology.
3. Showing that textual data is less predictive of ideology than interactions data.

The first two dot points have important policy upshots, though we only discuss the first point for the sake of brevity. We do not believe the third dot point should inform policy as we believe this finding is due to an idiosyncratic feature of our data as detailed in Section 5.1.

6.1.1 Digital footprints predict complex ideology

As noted in Section 3.5, most studies in this area use a simple, binary variable to represent an individual's ideology (i.e. left/right, Democrat/Republican, etc.).

Our results show that digital footprints can predict ideology in a nine-class problem where ideology is a complex variable that contains both an economic and social element.

We also showed that we can predict economic ideology above baseline accuracy when we account for the existence of centrists, but our predictions are far better when we omit centrists. In general, the more classes we include, the poorer our classifiers performed.

Consider a (for the sake of discussion, right wing party's) election strategist with access to digital footprints from a broad range of voters as well as a model that maps these footprints to ideology. This strategist would likely want to know as much about each voter's ideology as possible. They might wish to emphasise their party's candidate's libertarian economic policy to the 'libertarian-right' class voters whilst emphasizing their candidate's 'tough on crime' policies to 'authoritarian right' voters and targeting economically left leaning voters with advertisement designed to discourage them to vote. That is to say, they want a model that can accurately assign each voter to one of many ideological classes. Our results indicate that such a model might be hard to develop.

In contrast, an authoritarian government may want a model that can differentiate the majority of those with conforming ideologies from those with dissenting views. Our results show that this is a much more plausible task; models work best when separating two, relatively opposed ideologies with the absence of 'noisy' ideologies, i.e. those occupying the middle-ground between the two extremes.

Consequently, our results show that particular types of abuses of digital records are more feasible than others, which has important upshots for policy makers working in the digital privacy space.

6.2 Reddit's role in social data science

Our paper has illustrated that data from Reddit can be fruitfully utilized in the social sciences. As mentioned, most prior studies in this area relied on Facebook data and linked digital footprints such as Facebook Likes with private traits. The fact that Facebook data is no longer easily accessible to researchers is naturally a poor outcome for this area of inquiry. However, by illustrating Reddit data to be a viable alternative for this sort of inquiry, we hope to revitalize this field.

For instance, there are a broad range of Reddit communities where users reveal certain traits (whether through flairs or other means) that can be studied to examine the links between traits and online behaviour, or traits and behaviour full stop (using online actions as a proxy for general behaviour, i.e. posting in drug related subreddits can be considered a proxy for drug use).

Further, the data available from Reddit may be useful in other sorts of social science inquires. For instance, using our data set, we could examine the link between flaired ideology and participation in misogynistic online communities. These kinds of insights into the links between all sorts of social behaviours could inspire theories and experiments in a range of disciplines such as psychology, sociology and economics, and drive progress there too.

6.3 Limitations

We aimed to illustrate that digital records predict ideology and have shown this to be the case with the data we have collected from Reddit. However, it is unclear how well our results generalize. Our data comes from an online community for which the range of ages, interests and personality traits is likely narrower than in the general population. This issue is discussed extensively in Section 3.4 so we do not repeat this exposition here for the sake of brevity.

Conversely, our findings may underestimate the true extent to which digital footprints can predict ideology. The kinds of entities that are capable of misusing digital data in the ways we have discussed are governments and massive technology companies. Governments with sufficient surveillance capacity could, in principle, have access to every single digital footprint (Google searches, Facebook messages, etc.) for each resident in their country from which they could develop powerful models. Tech companies also have a huge wealth of data on their users. For example, Facebook has 2.85 billion monthly active users[42] with records of all their Likes, the contents of their messages, how long they wait before replying to certain people, etc. In contrast, we are constrained to using publicly available information which we can scrape and manage on standard personal computers. Consequently, our results are perhaps an underestimate of the capabilities of the kinds of entities liable to estimate ideology from digital footprints with malignant intentions.

Given more time and computational resources it would be interesting to gather a larger sample and employ a wider array of learning methods. We discuss this more in the next section.

6.4 Further research

There is substantial scope for further research to develop an even clearer picture of privacy risks imposed by the availability of huge digital records.

In this paper we used relatively simple predictive methods to map footprints to ideology. We showed that digital footprints can predict ideology but it is very possible that digital footprints can predict ideology with far greater accuracy than we have demonstrated in this paper.

This is a complex modelling problem with many problems (sparsity, class imbalance, high dimensional) and candidate solutions. For instance, we used a singular value decomposition to reduce the dimensions of our data but did not experiment with other methods of dimension reduction, i.e. UMAP[43]. Further, though our study uses a broader array of supervised learning algorithms than Kosinski et al.'s seminal paper[3] it could be rewarding to experiment with an even broader range of learning models, factorization machines in particular look like a promising algorithm for this problem. Likewise, there are many other methods of extracting features from textual data that may be more easily mapped to ideology.

In a similar vein, we could also extract different or additional features from our user-history records. In our user-interaction matrix, we extracted the number of times each user commented or posted in a particular subreddit. However, we could also extract the average (or total, maximum, minimum, etc.) score ('karma' in Reddit vernacular) of the user's posts/comments in each subreddit. Due to time constraints, we did not experiment with such an approach but these features could plausibly be used in conjunction with our existing user-interaction matrix to better predict ideology.

Additionally, further contributions to this area could be made through a similar study with an expanded sample. Our sample size is larger than that of Kosinski et al.'s[3], which is the most methodologically similar paper to our own. However, we were still limited by time and resources. Our data was scraped over the course of several weeks using personal computers. With sufficient knowledge of scraping (to optimize the process) and greater resources and time, it would be possible to construct a much larger sample. It would be easier for learning algorithms to differentiate signal from noise in a larger sample which would be conducive to further refining our understanding of digital footprints' predictive power over ideology. Similarly, with sufficient storage and computational power to process more data, we could have collected more than (a maximum of) 100 comments per user.

Appendix A

Scripts

This appendix outlines the psuedo code for the user flair scraper, user history scraper, data manipulator, comment scraper and comment manipulator scripts.

Appendix B

Recoding

In the raw user flair data, there are 12 unique flairs recorded: ‘:CENTG: - Centrist’, ‘:centrist: - Centrist’, ‘:centrist: - Grand Inquisitor’, ‘:left: - Left’, ‘:libright: - LibRight’, ‘:libright2: - LibRight’, ‘:right: - Right’, ‘:libleft: - LibLeft’, ‘:lib: - LibCenter’, ‘:auth: - AuthCenter’, ‘:authleft: - AuthLeft’, ‘:authright: - AuthRight’.

Clearly, some of these are duplicates so we map each of the raw flairs to one of nine flairs like so:

- ‘:CENTG: - Centrist’ → ‘centrist’
- ‘:centrist: - Centrist’ → ‘centrist’
- ‘:centrist: - Grand Inquisitor’ → ‘centrist’
- ‘:left: - Left’ → ‘left’
- ‘:libright: - LibRight’ → ‘libright’
- ‘:libright2: - LibRight’ → ‘libright’
- ‘:right: - Right’ → ‘right’
- ‘:libleft: - LibLeft’ → ‘libleft’
- ‘:lib: - LibCenter’ → ‘libcenter’
- ‘:auth: - AuthCenter’ → ‘authcenter’
- ‘:authleft: - AuthLeft’ → ‘authleft’
- ‘:authright: - AuthRight’ → ‘authright’

We use the resulting nine classes as our dependent variable in models for the nine class problem.

For the economic problem, we map the nine classes to the one of three classes like so:

- ‘centrist’ → ‘center’
- ‘left’ → ‘left’
- ‘libright’ → ‘right’
- ‘right’ → ‘right’
- ‘libleft’ → ‘left’
- ‘libcenter’ → ‘center’
- ‘authcenter’ → ‘center’
- ‘authleft’ → ‘left’
- ‘authright’ → ‘right’

For the social problem:

- ‘centrist’ → ‘center’
- ‘left’ → ‘center’
- ‘libright’ → ‘lib’
- ‘right’ → ‘center’
- ‘libleft’ → ‘lib’
- ‘libcenter’ → ‘lib’
- ‘authcenter’ → ‘auth’
- ‘authleft’ → ‘auth’
- ‘authright’ → ‘auth’

Appendix C

Model details

Specific details of all models used

Appendix D

Subreddits removed

We removed the following subreddits from our user-interaction matrix:

- r/Libertarian
- r/Anarchism
- r/socialism
- r/progressive
- r/Conservative
- r/democrats
- r/Liberal
- r/Republican
- r/Liberty
- r/Labour
- r/Marxism
- r/Capitalism
- r/Anarchist
- r/republicans
- r/conservatives

Appendix E

SVD results

Figure E.1 and Figure E.2 show users from the training and validation set in SVD space for SVD components ranging from 1-18 to illustrate the 'random' scattering of centrist users.

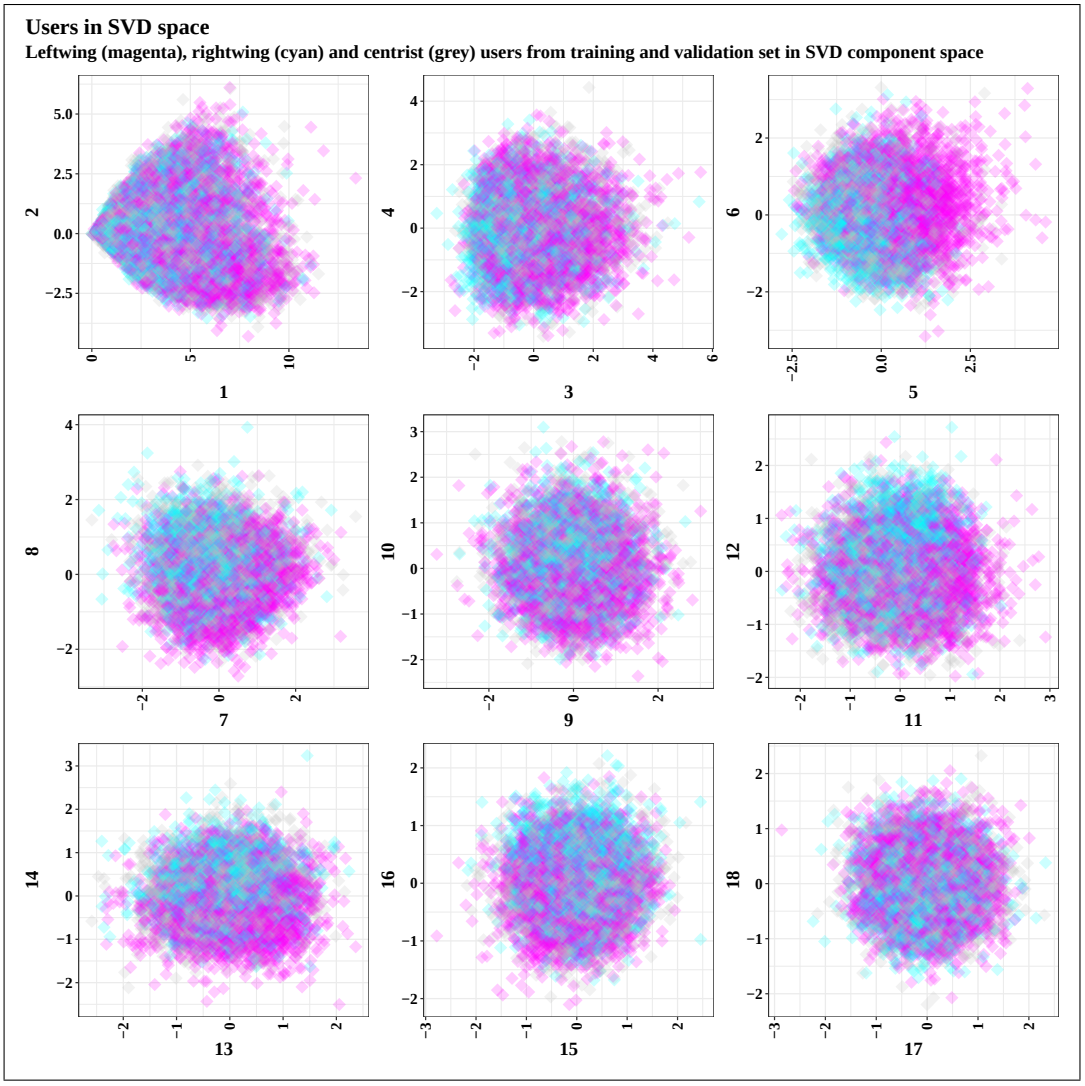
Users in SVD space
Leftwing (magenta), rightwing (cyan) and centrist (grey) users from training and validation set in SVD component space
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

FIGURE E.1

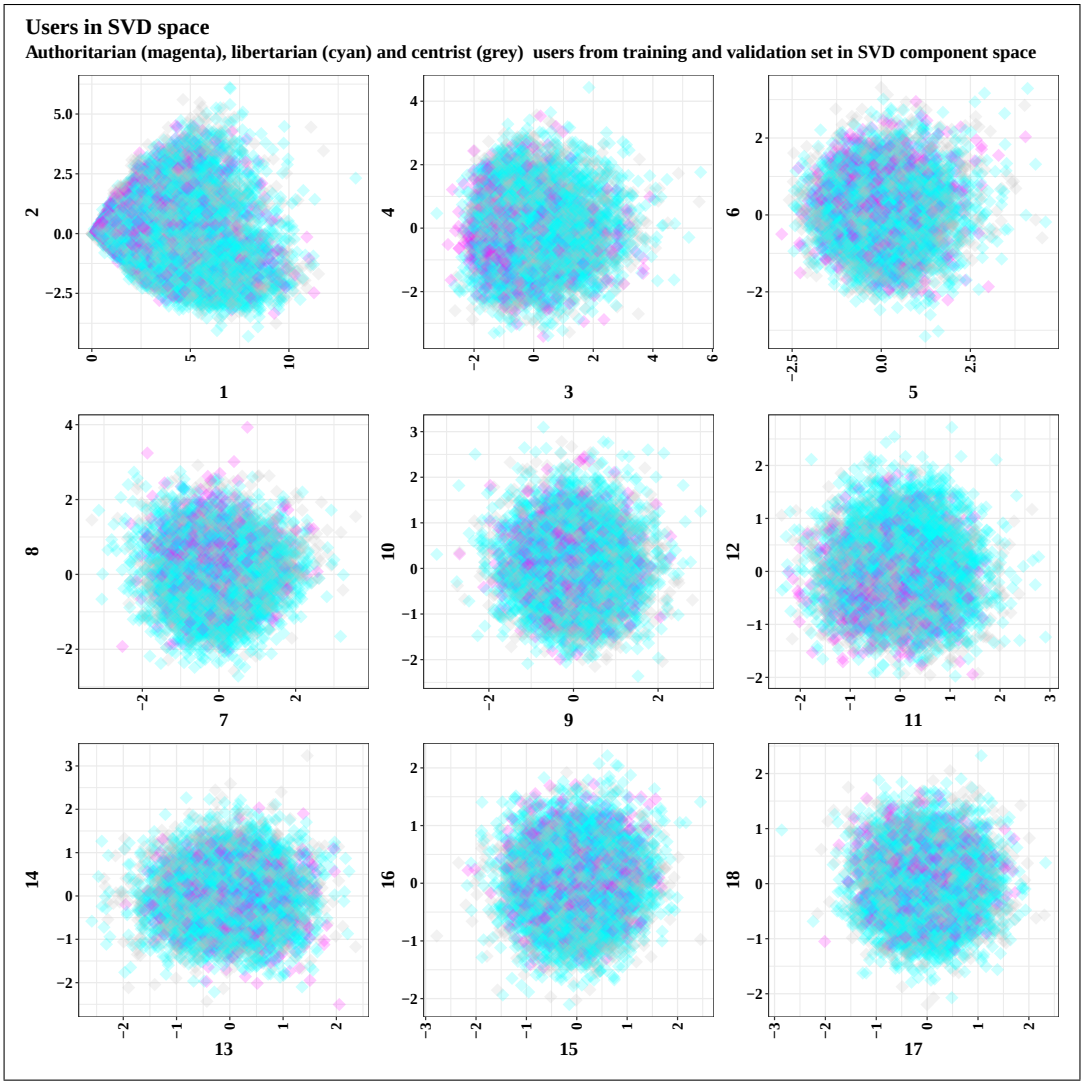
Users in SVD space
Authoritarian (magenta), libertarian (cyan) and centrist (grey) users from training and validation set in SVD component space
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

FIGURE E.2

Bibliography

[1] Emma Graham-Harrison, Carole Cadwalladr, and Hilary Osborne. Cambridge Analytica boasts of dirty tricks to swing elections. *The Guardian*, 2018. URL https://www.theguardian.com/uk-news/2018/mar/19/cambridge-analytica-execs-boast-dirty-tricks-honey-traps-elections.

[2] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246, 2018.

[3] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805, 2013.

[4] Michele Settanni, Danny Azucar, and Davide Marengo. Predicting individual characteristics from digital traces on social media: A meta-analysis. *Cyberpsychology, Behavior, and Social Networking*, 21(4):217–228, 2018.

[5] C Anderson and D Durbin. *One Person, No Vote: How Voter Suppression Is Destroying Our Democracy*. Bloomsbury Publishing, 2018. ISBN 9781635571387. URL https://books.google.com.au/books?id=KLlHDwAAQBAJ.

[6] Uwe Wolfradt and Jean E Pretz. Individual differences in creativity: personality, story writing, and hobbies. *European Journal of Personality*, 15(4):297–310, jul 2001. ISSN 0890-2070. doi: 10.1002/per.409. URL https://doi.org/10.1002/per.409.

[7] Robert R McCrae and Paul T Costa Jr. Conceptions and correlates of openness to experience., 1997.

[8] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E P Seligman, and Lyle H Ungar. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE*, 8 (9):e73791, sep 2013. URL https://doi.org/10.1371/journal.pone.0073791.

[9] Craig Ross, Emily S Orr, Mia Sisic, Jaime M Arseneault, Mary G Simmering, and R Robert Orr. Personality and motivations associated with Facebook use. *Computers in Human Behavior*, 25(2):578–586, 2009. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2008.12.024. URL https://www.sciencedirect.com/science/article/pii/S0747563208002355.

[10] Yair Amichai-Hamburger and Gideon Vinitzky. Social network use and personality. *Computers in Human Behavior*, 26(6):1289–1295, 2010. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2010.03.018. URL https://www.sciencedirect.com/science/article/pii/S0747563210000580.

[11] Jennifer Golbeck, Cristina Robles, and Karen Turner. *Predicting personality with social media*. jan 2011. doi: 10.1145/1979742.1979614.

[12] Samuel D Gosling, Adam A Augustine, Simine Vazire, Nicholas Holtzman, and Sam Gaddis. Manifestations of personality in Online Social Networks: self-reported Facebook-related behaviors and observable profile information. *Cyberpsychology, behavior and social networking*, 14(9):483–488, sep 2011. ISSN 2152-2723. doi: 10.1089/cyber.2010.0087. URL https://pubmed.ncbi.nlm.nih.gov/21254929https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3180765/.

[13] Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. Personality and Patterns of Facebook Usage. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 24–32, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312288. doi: 10.1145/2380718.2380722. URL https://doi.org/10.1145/2380718.2380722.

[14] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

[15] J Golbeck, C Robles, M Edmondson, and K Turner. Predicting Personality from Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 149–156, 2011. ISBN VO -. doi: 10.1109/PASSAT/SocialCom.2011.33.

[16] D Quercia, M Kosinski, D Stillwell, and J Crowcroft. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 180–185, 2011. ISBN VO -. doi: 10.1109/PASSAT/SocialCom.2011.26.

[17] Danny Azucar, Davide Marengo, and Michele Settanni. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, 124:150–159, 2018. ISSN 0191-8869. doi: https://doi.org/10.1016/j.paid.2017.12.018. URL https://www.sciencedirect.com/science/article/pii/S0191886917307328.

[18] Tal Yarkoni. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373, jun 2010. ISSN 00926566. doi: 10.1016/j.jrp.2010.04.001.

[19] C Segalin, A Perina, M Cristani, and A Vinciarelli. The Pictures We Like Are Our Image: Continuous Mapping of Favorite Pictures into Self-Assessed and Attributed Personality Traits. *IEEE Transactions on Affective Computing*, 8(2):268–285, 2017. ISSN 1949-3045 VO - 8. doi: 10.1109/TAFFC.2016.2516994.

[20] Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D Gosling, Gabriella M Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, and Others. Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30): 17680–17687, 2020.

[21] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48):12714–12719, nov 2017. ISSN 0027-8424. doi: 10.1073/PNAS.1710966114. URL https://www.pnas.org/content/114/48/12714.

[22] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication*, 64(2):317–332, 2014.

[23] M D Conover, B Goncalves, J Ratkiewicz, A Flammini, and F Menczer. 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*, pages 192–199, 2011. ISBN VO -. doi: 10.1109/PASSAT/SocialCom.2011.34.

[24] Daniel Preotiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. *Beyond Binary Labels: Political Ideology Prediction of Twitter Users*. jan 2017. doi: 10.18653/v1/P17-1068.

[25] Michele Vecchione, Harald Schoen, José Luis González Castro, Jan Cieciuch, Vassilis Pavlopoulos, and Gian Vittorio Caprara. Personality correlates of party preference: The Big Five in five big European countries. *Personality and Individual Differences*, 51(6):737–742, oct 2011. ISSN 01918869. doi: 10.1016/j.paid.2011.06.015.

[26] Antonio Chirumbolo and Luigi Leone. Personality and politics: The role of the HEXACO model of personality in predicting ideology and voting. *Personality and Individual Differences*, 49(1):43–48, jul 2010. ISSN 01918869. doi: 10.1016/j.paid.2010.03.004.

[27] Alain Van Hiel and Ivan Mervielde. Openness to Experience and Boundaries in the Mind: Relationships with Cultural and Economic Conservative Beliefs. *Journal of Personality*, 72(4):659–686, aug 2004. ISSN 0022-3506. doi: https://doi.org/10.1111/j.0022-3506.2004.00276.x. URL https://doi.org/10.1111/j.0022-3506.2004.00276.x.

[28] John T Jost, Jack Glaser, Arie W Kruglanski, and Frank J Sulloway. Political conservatism as motivated social cognition., 2003.

[29] D W Brogan. Political Parties: Their Organization and Activity in the Modern State. By Maurice Duverger. Translated by Barbara and Robert North. (New York: John Wiley & Sons, Inc.1954. Pp. xxxvii, 439.). *American Political Science Review*, 49(3):889–890, 1955. ISSN 0003-0554. doi: DOI: 10.1017/S0003055400296706. URL https://www.cambridge.org/core/article/political-parties-their-organization-and-activity-in-the-modern-state-by-maurice-A1A2BCA6E02F514D753477284DA6238C.

[30] Toni Rodon. Do All Roads Lead to the Center? The Unresolved Dilemma of Centrist Self-Placement. *International Journal of Public Opinion Research*, 27(2):177–196, jun 2015. ISSN 0954-2892. doi: 10.1093/ijpor/edu028. URL https://doi.org/10.1093/ijpor/edu028.

[31] M Kosinski, Yilun Wang, Himabindu Lakkaraju, and J Leskovec. Mining big data to extract patterns and predict real-life outcomes. *Psychological methods*, 21 4:493–506, 2016.

[32] F Pedregosa, 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. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. jan 2013. URL https://arxiv.org/abs/1301.3781.

[34] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. oct 2013. URL https://arxiv.org/abs/1310.4546.

[35] Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features BT - Machine Learning: ECML-98. pages 137–142, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg. ISBN 978-3-540-69781-7.

[36] T Fawcett. Using rule sets to maximize ROC performance. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 131–138, 2001. doi: 10.1109/ICDM.2001.989510.

[37] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015. ISBN 1498712169.

[38] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370.

[39] Md Nasim Adnan and Md Islam. *One-Vs-All Binarization Technique in the Context of Random Forest*. jan 2015.

[40] Ji Zhu, Saharon Rosset, Hui Zou, and Trevor Hastie. Multi-class AdaBoost. *Statistics and its interface*, 2, feb 2006. doi: 10.4310/SII.2009.v2.n3.a8.

[41] Raúl Rojas. AdaBoost and the Super Bowl of Classifiers A Tutorial Introduction to Adaptive Boosting. 2009.

[42] Facebook. FORM 10-Q, Q1 2021. Technical report, Facebook, 2021.

[43] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. feb 2018. URL https://arxiv.org/abs/1802.03426.