Can digital footprints accurately predict political ideology? Evidence from Reddit

Michael Kitchener (32220162)
mkit0007@student.monash.edu

Supervisor: Assoc Prof. Paul Raschky
Paul.Raschky@monash.edu

May 6, 2021

1 Topic

This paper seeks to establish 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 Reddit users and will be recording their comments and posts in a range of 'subreddits' (interest based discussion forums) so that we can model their political ideology as a function of the extent to which they engage with different subreddits. With the rapid uptake of social media, there is a growing body of work focused on determining an individual's traits (personality traits, gender, sexuality, etc.) through their digital interactions. Typically, these have used textual data from Twitter or Facebook data which, owing to increased public scrutiny over online privacy and prominent incidents such as the Cambridge Analytica scandal, is no longer easily accessible. This paper aims to make a contribution to these efforts through utilizing an original data set from a neglected source of social data with a large sample size, a sophisticated measure of ideology and rich digital footprints. In doing so we aim to contribute to the literature on how, through online behaviour, people may implicitly disclose private information and thus inform policy on digital privacy.

2 Background and Motivation

The extent to which publicly available records of online activity can predict ideology has important real world implications. 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 seemingly innocuous 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 such as voter suppression, which some consider to be anti-democratic.

Consequently, we are motivated by a desire to illuminate the extent to which political ideology can be predicted from digital records. In doing so we contribute to a growing body of work illustrating how online behaviour can implicitly disclose private traits. It has beeen shown that neural networks can accurately predict sexual orientation from pictures of an individual's face (Wang and Kosinski, 2018) and that Facebook likes can accurately predict big 5 personality

traits (Kosinski et al., 2013). Further, 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 (Youyou et al., 2015). Other findings illustrate that mobile phone use data accurately predicts big 5 traits (Stachl et al., 2020). Results are best summarized in Settanni's 2018 meta-analysis: "digital traces from social media can be studied to assess and predict theoretically distant psychosocial characteristics with remarkable accuracy" (Settanni et al., 2018).

Some attention has been payed to political traits; Colleoni shows that natural language processing techniques can be used to accurately classify the political leanings of Twitter users (Colleoni et al., 2014), however, most work in this area is focused on predicting psychological traits. Since the extent to which online records can predict political leanings and behaviours is particularly salient with respect to the privacy concerns we have discussed, we are motivated to further illuminate this area.

Our core approach is taken from Kosinski's seminal work in this field (Kosinski et al., 2013) which illustrated that a range of personal traits (including ideology) can be predicted from someone's Facebook likes.

3 Contribution

We expand upon Kosinski's work in several ways:

Firstly, we are using a larger and likely richer data source. 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 disclose this information via their Facebook likes). These interests may be connected to ideology, leading to a stronger model and a better indication of the predictive power of digital data for personal traits. Further, Kosinski's Facebook data does not indicate the extent to which someone is interested in a particular topic (i.e. both a typical fan and an extreme fan of a particular band may indicate interest through liking the band's Facebook page; Facebook likes cannot illustrate varying degrees of interest). In contrast, we can model ideology on the extent to which an individual engages with a particular interest (through either the number of posts/comments they have made in certain subreddits, or the score of their posts in that subreddit) which may be informative with respect to ideology. Our data also features a more sophisticated response variable (multiple categories corresponding to a 2-dimensional conception of ideology) than Kosinski's, in which ideology is coded as a binary split. Finally, we will utilise more advanced predictive methods in order to more accurately reflect the extent to which digital data can predict ideology. Kosinski uses logistic regression on the PCA components of his user-like matrix to model ideology. We will try a variety of statistical learning methods. This should ultimately further our knowledge of the privacy risks implied by digital footprints and therefore the need for scrutiny of our online behaviour and for policy providing adequate protection of digital data.

We also seek to contribute to the field of 'social data science' through illustrating the viability of Reddit as a source of digital data that can be easily obtained and applied to a variety of research questions.

4 Methodology

4.1 Data

Data was scraped from the 'r/PoliticalCompassMemes' subreddit. In this forum users from a range of political persuasions post comedic images and videos of a political nature. The majority of users choose to 'flair' their posts and comments with their ideology according to their results in the popular Political Compass Test (https://www.politicalcompass.org/test). Users are placed into one of four groups according to their economic and social beliefs:

- {left, authoritarian} i.e. communists
- {right, authoritarian} i.e. traditional conservatives
- {left, libertarian} i.e. social democrats
- {right, libertarian} i.e. libertarians

Note: in practise there are more flair options than these four, i.e. some users flair themselves as simply 'left' or 'right' indicating ambivalence on the 'social' side of things.

We obtained a sample of 91,000 users 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 flair as a row in the data set. This resulted in a data set of 91,000 username and flair combinations. The data set looks like this:

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

The digital footprints of these users will also be scraped using a Python script. The script has been created and tested but we have not yet finished gathering the footprints of all users in our sample. Each post or comment for a user is stored as a row in a data set that looks like this:

username	interaction	title	body	score	time	subreddit
user1	post	GTA V …	what a…	43	10:32 2/1/21	r/gaming
user1	comment		Mozart's …	12	16:12 8/12/20	r/classicalmusic
user2	post	Today's …	One of …	-6	5:36 4/3/21	r/boxing
user2	post	The …	George is …	3	4:24 4/3/21	r/seinfeld
user2	comment		Which …	0	5:09 3/3/21	r/crypto
…	…	…	…	…	…	…

This can easily 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. For example:

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
…	…	…	…	…

These form a set of predictors and can be merged with the user-flair data. This allows us to model ideology on digital interactions. We will build models on data that represents the number of times a user has posted/commented in a subreddit, the average score of their posts/comments in a subreddit, the total score of their posts/comments and whether or not they post/comment in a subreddit.

In order to examine the extent to which ideology may inadvertently be revealed we will remove columns that pertain to interactions with explicitly political subreddits.

Descriptive statistics and visualisation for ideology are reported below:

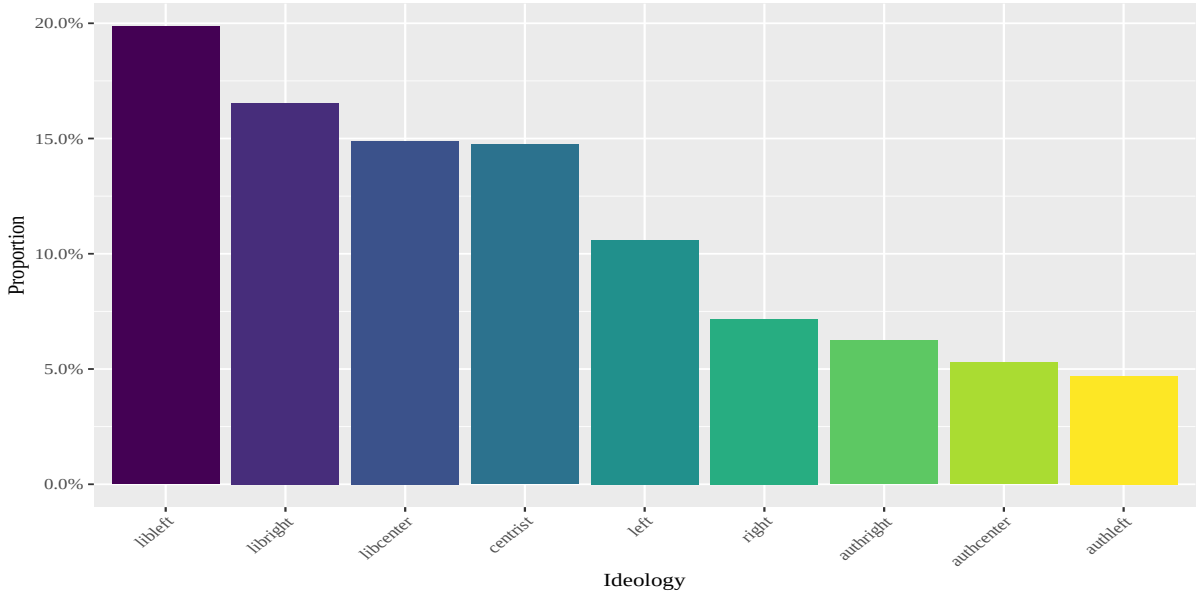
Table 1.1: Ideology frequency in sample (n = 91,000)

libleft	libright	libcenter	centrist	left	right	authright	authcenter	authleft
18,070	15,054	13,548	13,408	9,646	6,526	5,672	4,801	4,275

Table 1.2: Ideology proportion in sample (n = 91,000)

libleft	libright	libcenter	centrist	left	right	authright	authcenter	authleft
0.2	0.17	0.15	0.15	0.11	0.07	0.06	0.05	0.05

Fig 1.1: Ideology proportion in sample (n = 91,000)



4.2 Model

The final data set will undoubtedly:

1. Be very high dimensional

2. Be very sparse (a given user will not interact at all with the majority of subreddits comprising the columns)
3. Be unbalanced; there will not be even proportions of each possible flaired ideology in our sample

Due to these issues, the fact that we do not know the exact form of the data yet (number of predictor variables), and the fact that we will experiment with many different learning models, we have not specified any specific models here.

To deal with the first two issues we may use a subset of all our predictors, reduce dimensions via PCA or attach a lasso type penalty terms to models. Regarding the unbalanced nature of our response variable (ideology) we will experiment with under-sampling dominant ideologies, increasing the cost of mis-classifying minority ideology observations, and assess test set accuracy according to balanced accuracy.

We will trial many different models and report the differing levels of success. We will certainly look at a multinomial logit model, random forests and boosted tree algorithms. We will split the data into a training set to train the models and a test set to evaluate model performance.

5 Expected results

We expect to develop a reasonably strong predictive model. Time permitting we will also apply this predictive model to out of sample users and monitor how users of different predicted political ideologies' sentiment towards politicised topics has changed over time to illustrate the usefulness of Reddit data as a social sciences research tool.

6 Limitations

Limitations pertain to the generalisability of our results. We are using Reddit data and though we aim to contribute to the broader understanding of the extent to which personal traits may be implicitly revealed through online behaviour, it may be that this possibility on Reddit doesn't translate to other mediums.

Further, it is possible that our results are not representative of Reddit users more broadly. Users who post in the Political Compass Memes subreddit may have different online habits to users who do not and so it is possible the model does not generalise to all Reddit users.

Lastly, insofar as our our model illuminates the determinants of ideology it should be noted that Reddit users are not a random sample of the broader population.

7 Timeline

- May:
  - Complete literature review
  - Have digital histories of all 91,000 users scraped and stored
- July

    - Have developed set of predictive models
- August
    - Write up a first draft of results
    - Continue to edit and revise
- October
    - Submit final draft

References

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

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

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

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., and Others (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30):17680–17687.

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

Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.