Converting the Twitter API into a High Quality Online Panel via Machine Learning

Roberto Cerina
roberto.cerina@nuffield.ox.ac.uk
Maastricht University & Nuffield College, Oxford University

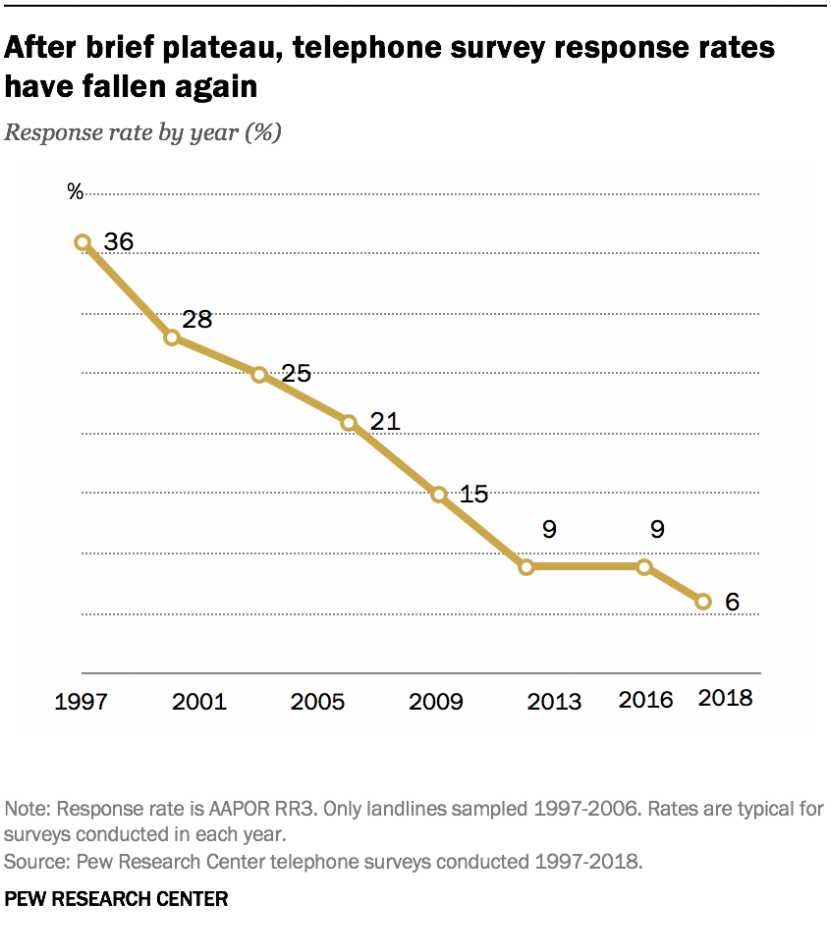
Kayvane Shakerifar
kayvane.shakerifar@gmail.com
Consulting Data Scientist

Raymond Duch
raymond.duch@nuffield.ox.ac.uk
Nuffield College, Oxford University

Introduction

Challenges to Traditional Polling Practices

i) Response Rates to Pollsters have plummeted[1]:



[1] https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/

ii) selection effects are severe and affect polling accuracy - e.g. education in 2016 election[Kennedy et al., 2018]:

Table 3. Share of college graduates in interactive voice response polls relative to the Current Population Survey in three states

Michigan		Pennsylvania		Wisconsin	
CPS benchmark	*38%*	*CPS benchmark*	*36%*	*CPS benchmark*	*37%*
Gravis	53%	Gravis	57%	Emerson	48%
Emerson College	48%	Emerson College	54%	Mitchell	N/A
Mitchell Research	N/A	Harper	54%	Trafalgar	N/A
Trafalgar Group	N/A	Trafalgar Group	N/A	PPP	N/A
EPIC/MRA	N/A	PPP	N/A		
PPP	N/A				

NOTE.—Benchmark data are weighted, filtered on self-reported voters, and come from the November 2016 Current Population Survey Voting and Registration Supplement. Election poll data come from pollster press releases and appear to be weighted. "N/A" indicates that respondent education level does not appear to have been measured in the poll.

Digital Trace: Opportunities & Challenges

- Online Partisan Behaviour directly reflects Real-life Partisan Behaviour (e.g. registering as R/D or attending a partisan primary)[Cerina and Duch, 2020];
- Social network reveals partisan identity[Barberá, 2015];
- Individual-level demographic information can also be deduced[Wang et al., 2019].

Opportunities:

- high frequency;
- low cost[2];
- rich in user characteristics.

Challenges:

i) **Unrepresentative** → Regularized Prediction & Post-Stratification[Gelman et al., 2013, Gelman, 2018];

ii) **Unstructured** → Machine Learning.

[2]Free to researchers, low-cost per observations for private sector

Machine Learning

Data

Data

Initial Twitter Corpus:

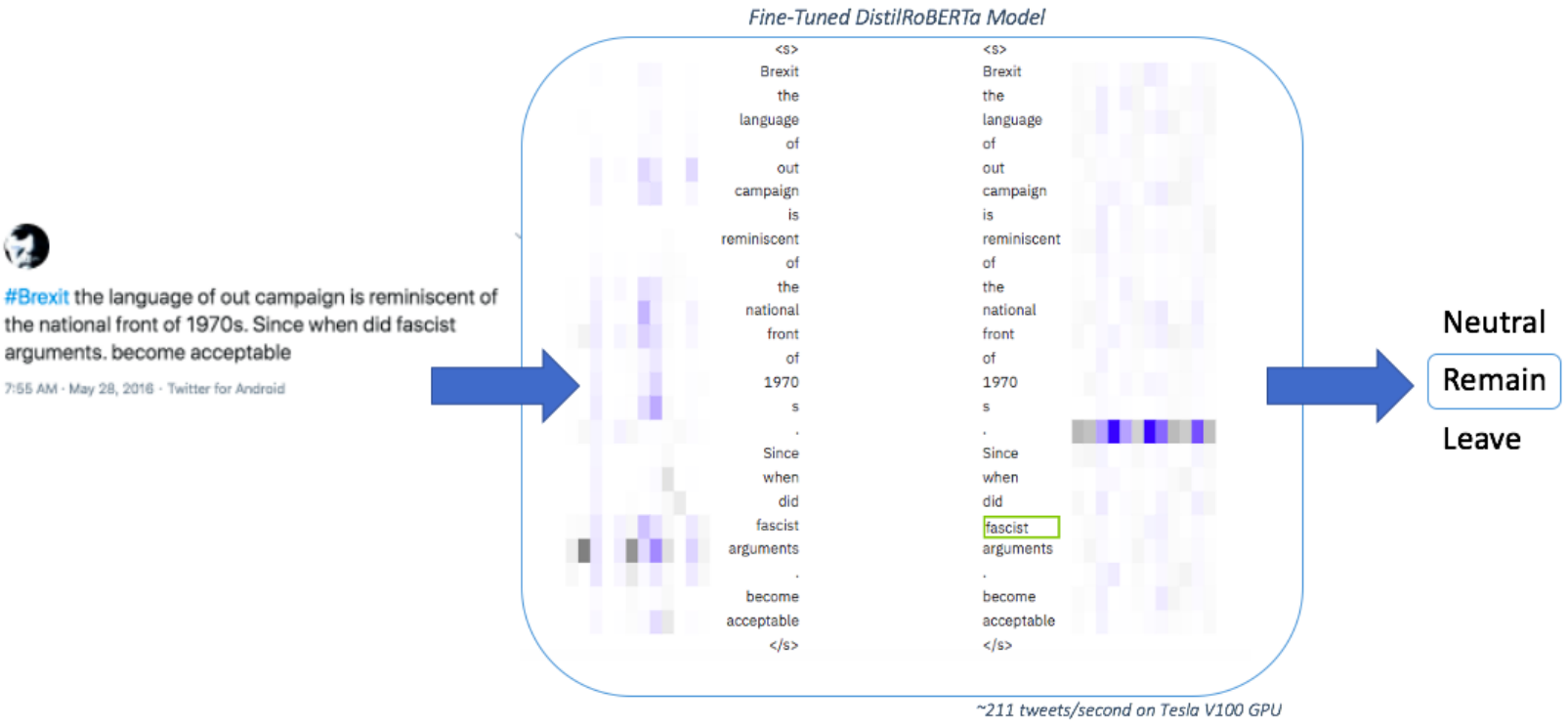
- 25 million Tweets;
- 3.8m users;
- over 9 weeks leading up to the Brexit vote.

Analyzed Twitter Corpus:

- $253,161$ users which could be geo-located. (API cost limitation)

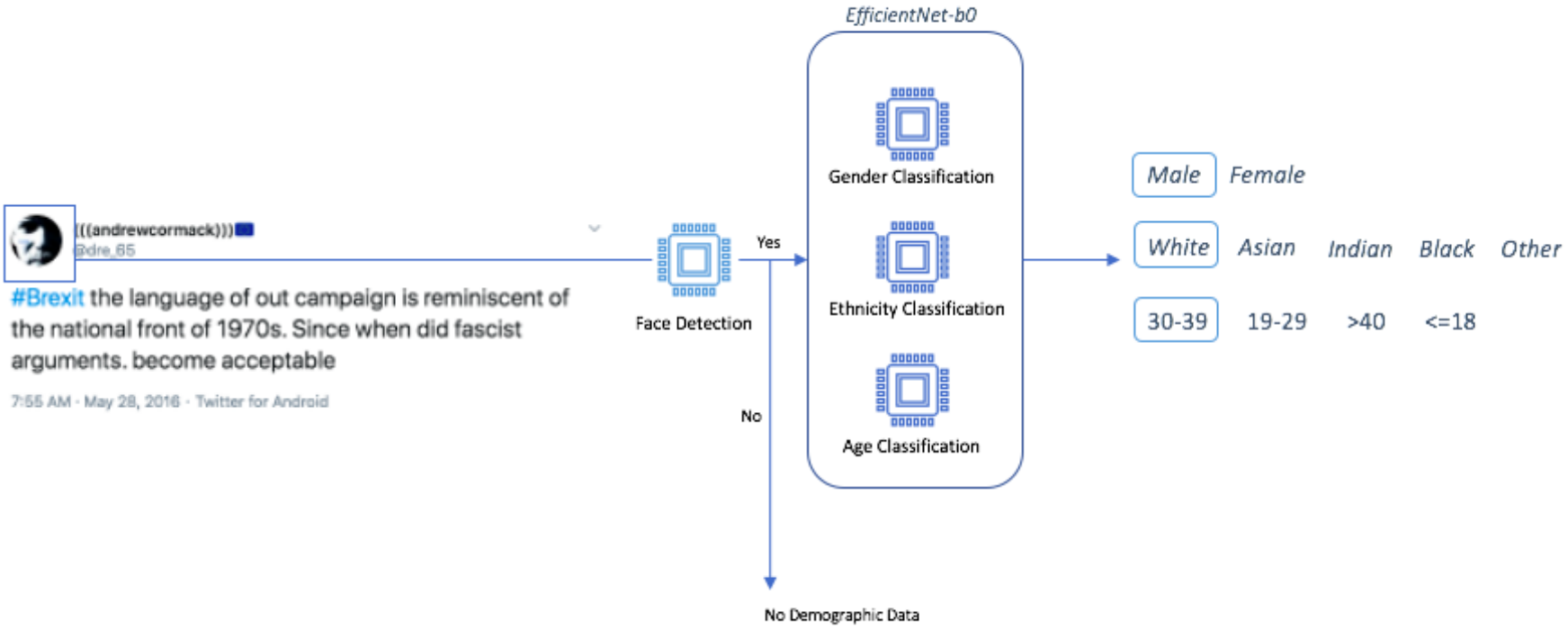
Feature Extraction

Brexit Stance



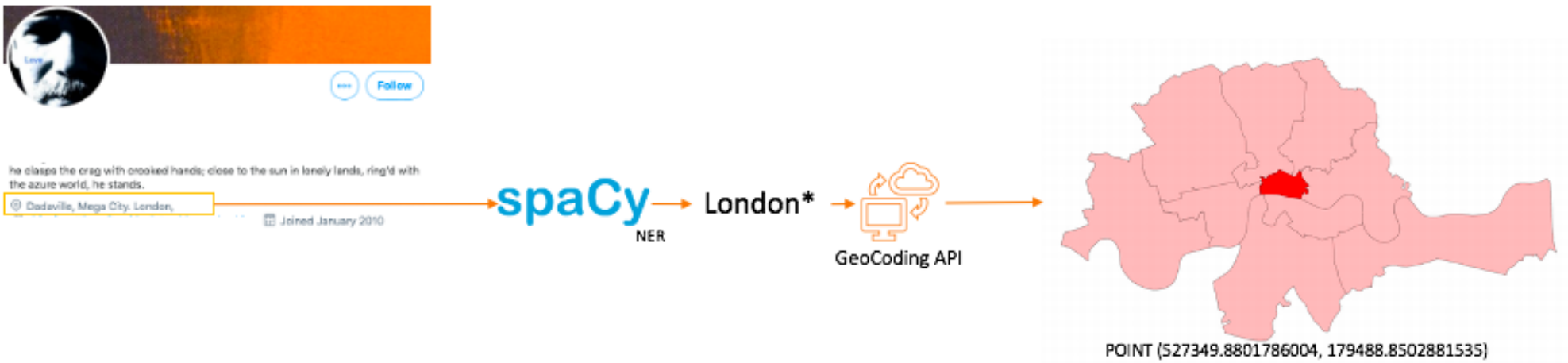
- Training Data → [Grčar et al., 2017];
- roBERTa → [Liu et al., 2019, Sanh et al., 2019]

Demographics



- Training Data → UTK-face[Zhang et al., 2017], ImageNet[Deng et al., 2009];
- Deep Residual Neural Network → [Tan and Le, 2019];

Location

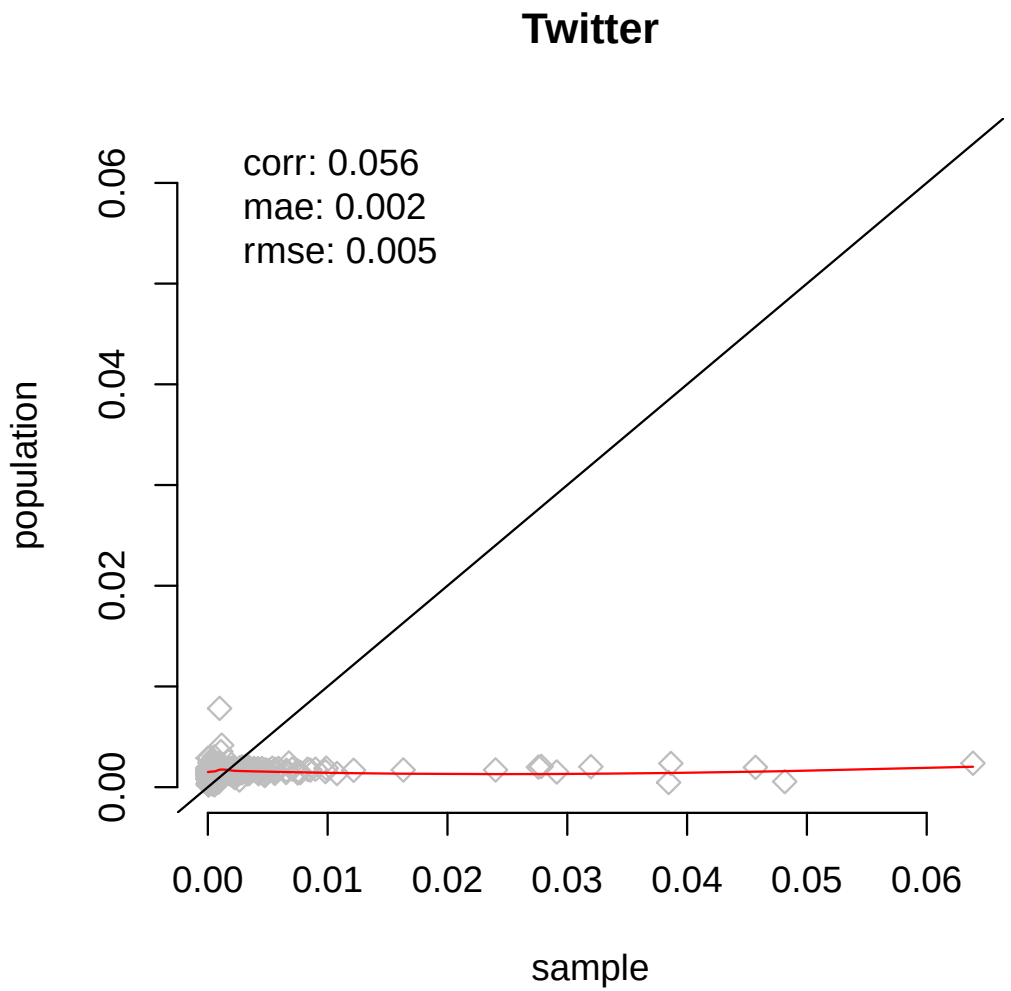


*For Illustrative Purposes, generic locations were discarded for downstream modelling forecasts

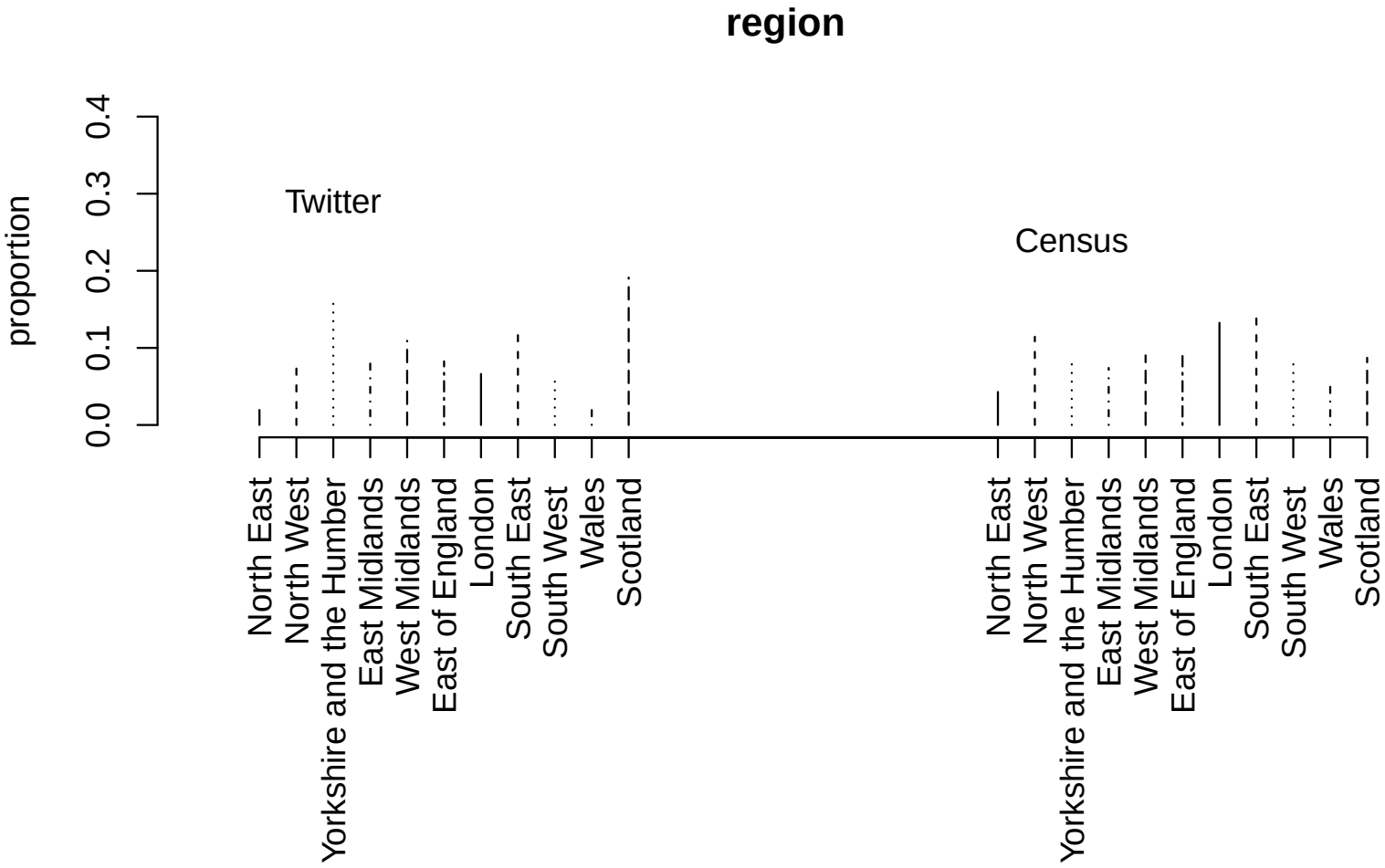
- Named Entity Recognition → spaCy [Honnibal and Montani, 2017];

Sample v. Population

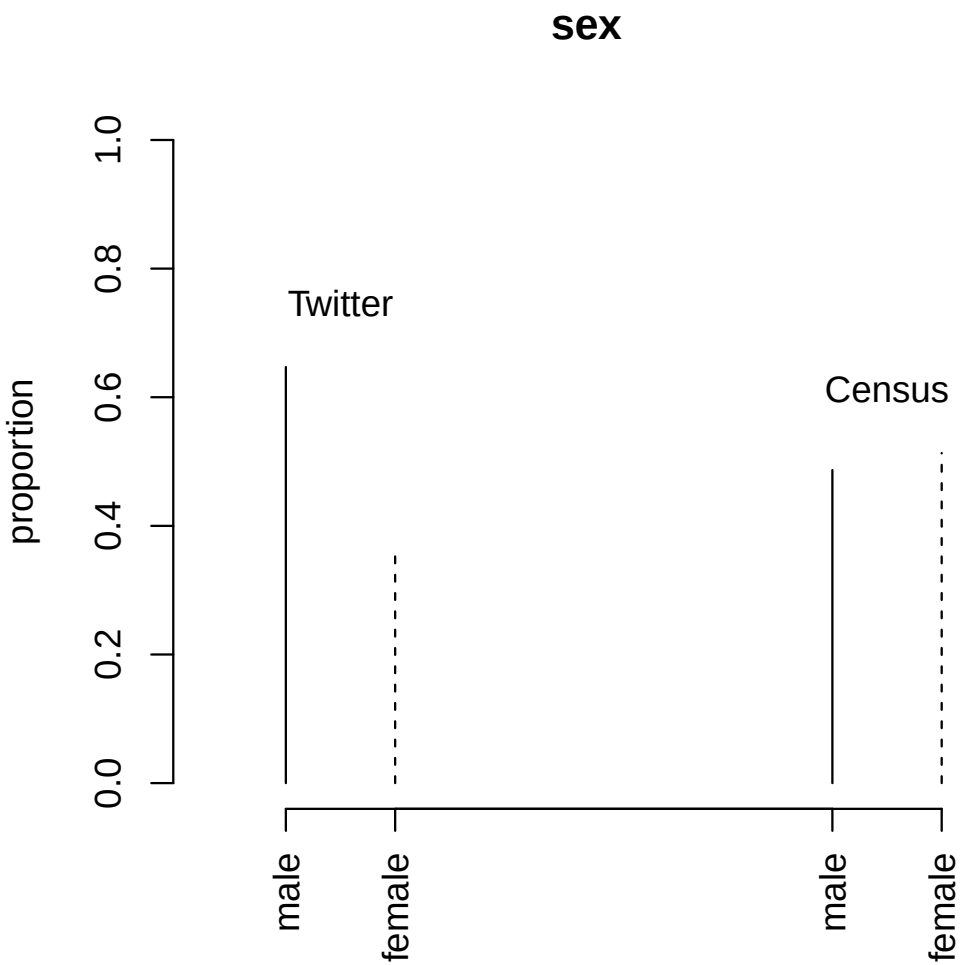
Sample v. Population: Constituency level



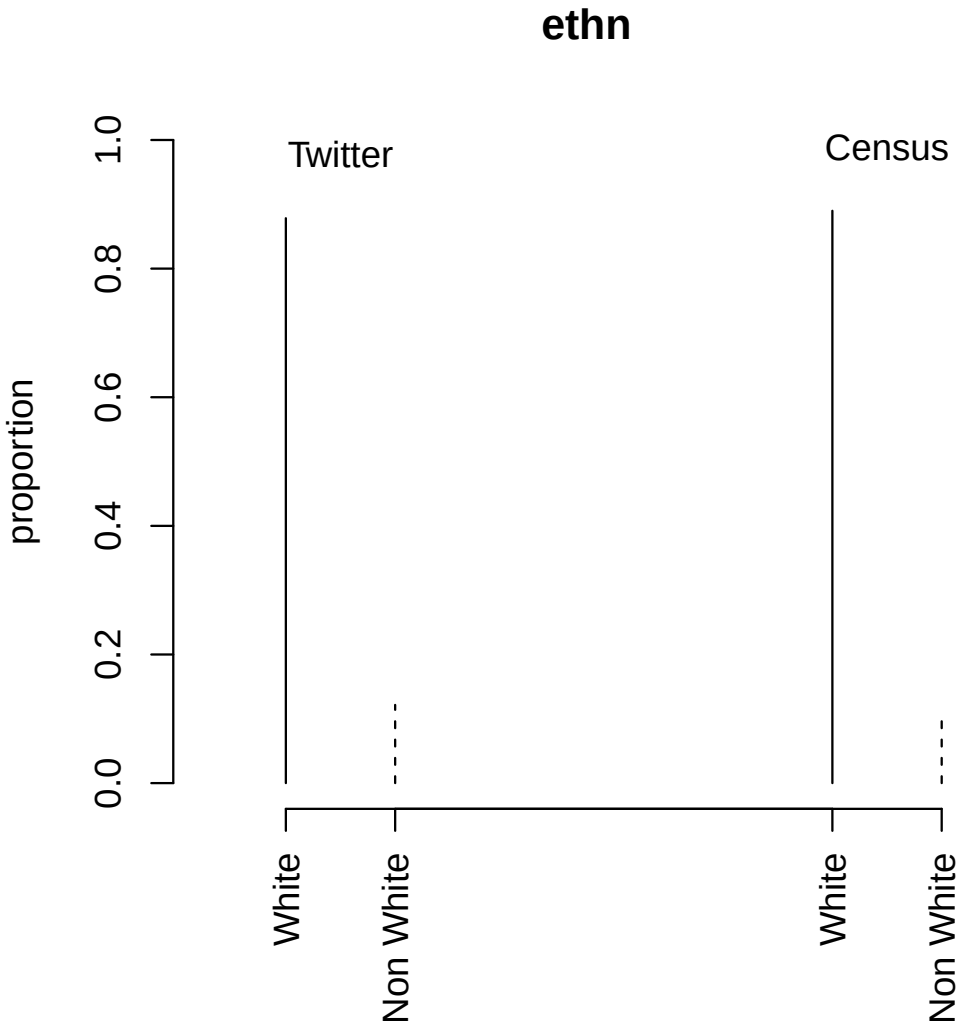
Sample v. Population: Region



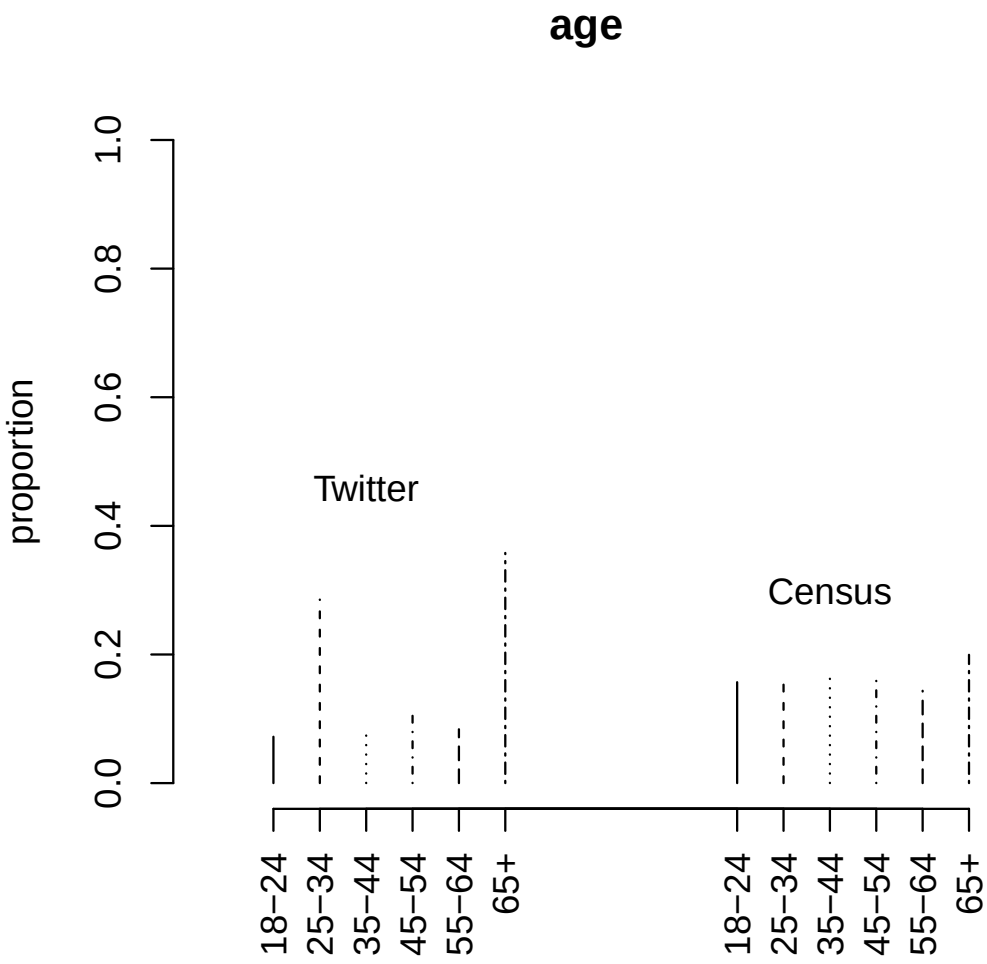
Sample v. Population: Sex



Sample v. Population: Ethnicity



Sample v. Population: Age



Predictions

i. collect individual-level sample → Twitter + British Election Study[Fieldhouse et al., 2016];

ii. prepare stratification frame with population-level counts - e.g. in constituency A there are $1,000$ individuals from the following *cell*: {Age:25 to 34; Ethnicity: Black; Sex: Male; Edu: Level 4; etc.} → Census Safeguarded Micro-Data[Longhurst et al., 2007];

iii. prepare a constituency-level predictor - e.g. constituency profile: {% Conservative GE2015 : 56%; % White: 95%; % in Agriculture: 50%; etc.};

iv. run a learner - Random Forest[Breiman, 2001] via `ranger`[Wright and Ziegler, 2015]:

$$\hat{\Pr}_i(\mathrm{T}=1) = \frac{1}{B}\sum_b \phi^T_{i,b}(x,\theta); \tag{1}$$

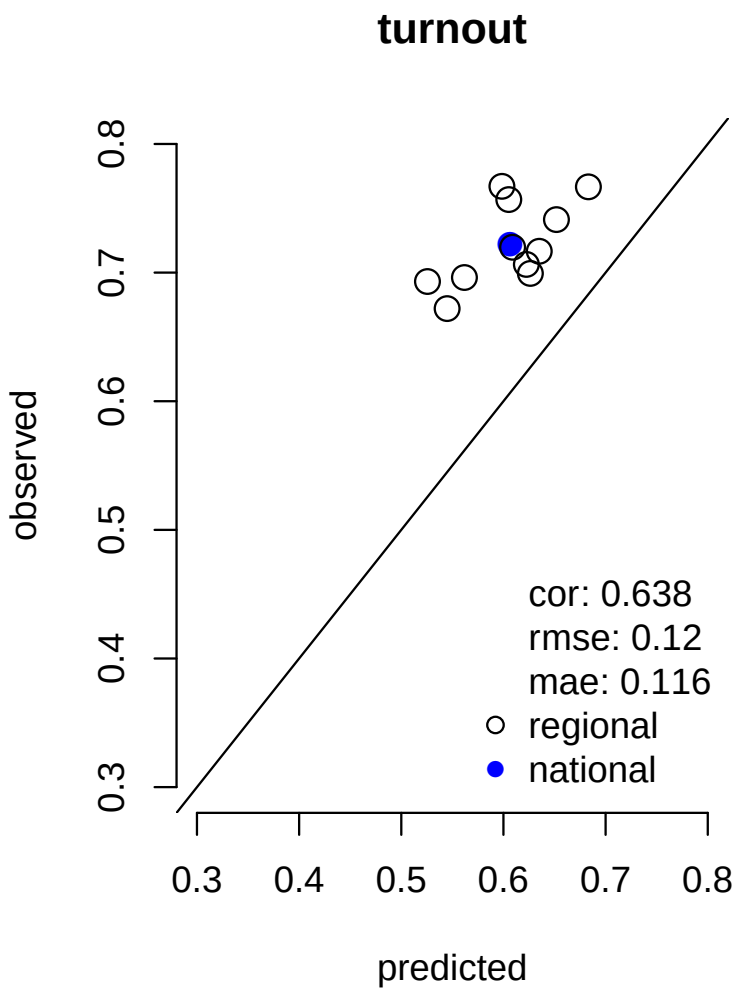
$$\hat{\Pr}_i(\mathrm{L}=1) = \frac{1}{B}\sum_b \phi^L_{i,b}(x,\theta \mid T); \tag{2}$$

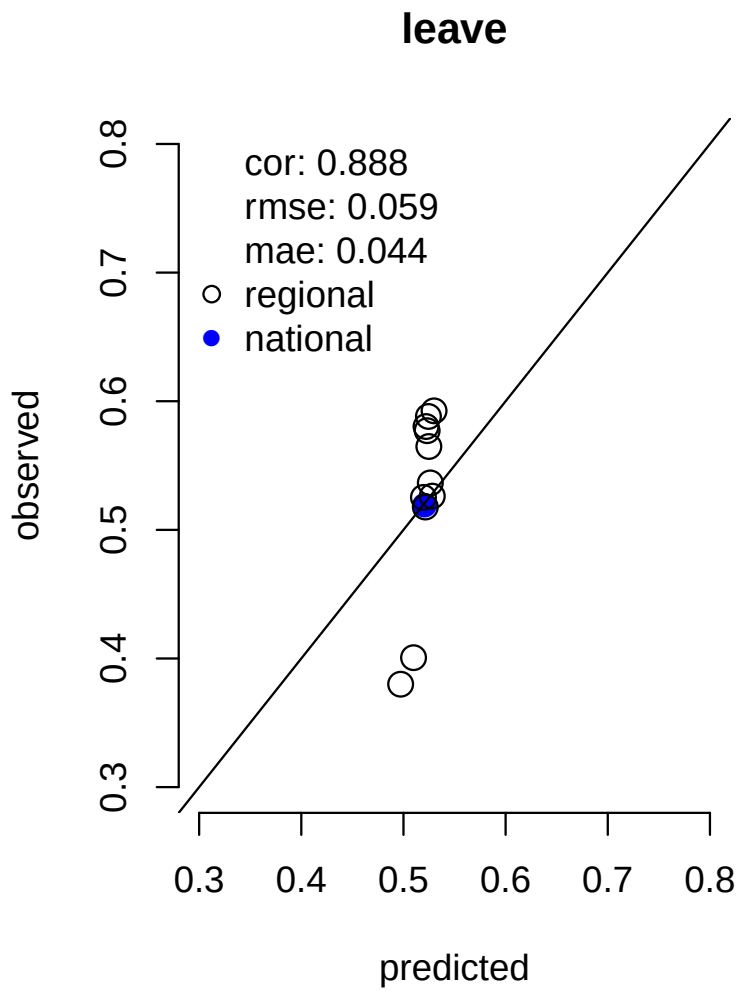
$$\tag{3}$$

v. aggregate to the area-level for each individual within the area $i \in a$.

$$\pi_a^{L=1} = \frac{\sum_{i \in a} \hat{\Pr}_i(\mathrm{L}=1) \times \hat{\Pr}_i(\mathrm{T}=1)}{\sum_{i \in a} \hat{\Pr}_i(\mathrm{T}=1)} \tag{4}$$

Leave Vote Share Predictions





Future Work

Future work

- Screen for Bots → `Botometer` API[Varol et al., 2017];
- Predict UK Party Identification[Barberá, 2015, Barberá and Rivero, 2015] + past vote;
- re-work age model → how to explain over-representation of 65+ ?;
- compare performance with traditional MrP - Variational Inference[Kucukelbir et al., 2015] for 200k observations;
- realistic estimates of uncertainty;
- constituency level results.

References I


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


Barberá, P. and Rivero, G. (2015).
Understanding the political representativeness of twitter users.
Social Science Computer Review, 33(6):712–729.


Breiman, L. (2001).
Random forests.
Machine learning, 45(1):5–32.


Cerina, R. and Duch, R. (2020).
Measuring public opinion via digital footprints.
International Journal of Forecasting.


Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009).
Imagenet: A large-scale hierarchical image database.
In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.


Fieldhouse, E., Evans, G., Green, J., Schmitt, H., v. d. E. C., Mellon, J., and Prosser, C. (2016).
British election study internet panel wave 8 (2016 eu referendum study, daily campaign survey) , doi: 10.15127/1.293723.

References III


Gelman, A. (2018).
Regularized prediction and poststratification (the generalization of mister p).
URL: https://statmodeling. stat. columbia. edu/2018/05/19/regularized-predictionpoststratification-generalization-mister-p.


Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013).
Bayesian data analysis.
CRC press.


Grčar, M., Cherepnalkoski, D., Mozetič, I., and Novak, P. K. (2017).
Stance and influence of twitter users regarding the brexit referendum.
Computational social networks, 4(1):6.


Honnibal, M. and Montani, I. (2017).
spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.
To appear, 7(1).


Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D., et al. (2018).
An evaluation of the 2016 election polls in the united states.

Public Opinion Quarterly, 82(1):1–33.


Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. (2015).
Automatic variational inference in stan.
In *Advances in neural information processing systems*, pages 568–576.


Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019).
Roberta: A robustly optimized bert pretraining approach.
arXiv preprint arXiv:1907.11692.


Longhurst, J., Tromans, N., Young, C., and Miller, C. (2007).
Statistical disclosure control for the 2011 uk census.
In *Joint UNECE/Eurostat conference on Statistical Disclosure Control, Manchester*, pages 17–19. Citeseer.


Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019).
Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.
arXiv preprint arXiv:1910.01108.

Tan, M. and Le, Q. V. (2019).
Efficientnet: Rethinking model scaling for convolutional neural networks.
arXiv preprint arXiv:1905.11946.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and Flammini, A. (2017).
Online human-bot interactions: Detection, estimation, and characterization.
arXiv preprint arXiv:1703.03107.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flock, F., and Jurgens, D. (2019).
Demographic inference and representative population estimates from multilingual social media data.
In *The World Wide Web Conference*, pages 2056–2067.


Wright, M. N. and Ziegler, A. (2015).
ranger: A fast implementation of random forests for high dimensional data in c++ and r.
arXiv preprint arXiv:1508.04409.


Zhang, Z., Song, Y., and Qi, H. (2017).
Age progression/regression by conditional adversarial autoencoder.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818.