Baselines for Demographic Inference on a New Gold Standard Twitter Corpus

Jason Radford, Luke Horgan, David Lazer
Network Science Institute
Northeastern University
Boston, MA 02131
contact: j.radford@northeastern.edu

***Abstract*—A variety of studies have shown that machine learning methods like convolutional neural nets and random forests can be used to accurately infer characteristics of people online such as their gender, age, race, or political orientation. However, these studies are based on labels generated using the data themselves, typically human coding of subjects, and presume subjects are authentic humans. This creates systematic selection biases owing what features humans can draw inferences from. In this preliminary study, we connect Twitter Data to an exogenous data source, public voter data, to create a new gold standard data set for inferring demographic information about online participants. We run a standard battery of machine learning algorithms on bag-of-words representations of individuals' twitter posts to generate new baselines for how well these characteristics can be predicted. Our baselines are substantially lower than most reported studies, suggesting sampling bias has led to an over-estimation of how well machine learning algorithms perform on this task.**

***Keywords*-computational social science, machine learning, twitter, demographic inference**

I. Introduction

A wide variety of research demonstrates that large, unstructured data from social media and the web can be used to make a variety of accurate inferences about social phenomena [1]. One particular subfield is demographic inference wherein big data is used to infer demographic characteristics about the person or people generating the data [2]. For example, [3] uses follower networks on Twitter to infer an individual's political leanings. [4] use Facebook posts to infer individuals' personality traits, age, and gender. In general, the robustness of these studies suggests that it may be difficult to find demographic characteristics that cannot be inferred through some form of everyday big data.

One of the principle obstacles to performing demographic inference using the vast troves of data is the lack of any sort of dependable training set. Twitter profiles are notoriously unreliable and often thoroughly incomplete. Is Jane Doe from Springfield this 36-year old who voted republican in the last presidential election, that 53-year old who voted democrat, some college student who didn't vote at all, or a bot which couldn't if it wanted to?

Previous research relies on hand-coding as a workaround. Human beings spend hours pouring over hundreds or thousands of Twitter accounts, manually attempting to identify demographic markers in photos, biographies, and tweets themselves to create a ground truth datasets. This process creates a range of biases in the kinds of people and characteristics we try to predict. In this study, we take advantage of a new panel of nearly 1.6 million Twitter accounts which have been linked with secondary data outside of Twitter: voter records.

II. Methods

In this study, we look to perform baseline calculations for the predictability of these standard demographic features to determine how accurate our predictions are and what characteristics are more easily predicted. For this baseline, we take a simple bag of words approach to feature selection and run a standard battery of predictive algorithms, including regularized regressions, support vector machines, and random forests, on each of the demographic characteristics.

III. Results

First, we find that race and gender are easier to predict than political party, age, and location. We posit that this is because of the societal salience they carry as master statuses. In other words, these are highly visible traits which feature prominently in people's words.

Second, we find that our algorithmic performance diverges substantially from that of published studies. This points to a fundamental flaw in the traditional reliance on hand-coded datasets. The sampling bias caused by hand-coded data appear to cause us to oversample accounts which are heavily gendered or politically and geographically revealing. Moreover, hand-coded datasets have practical limits on their size. The sheer quantity of accounts that we have at our disposal affords us relative flexibility in our pipeline; our results are robust to choice of predictive algorithm.

IV. Discussion

The use of big data for social analysis must rely on higher quality ground truth data that not only distinguishes human from bot behavior in these socio-technical systems but is gather exogenously from the data source. We cannot choose to include or exclude data that is easy to classify based solely on that data.

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

References

[1] D. Lazer and J. Radford, “Data ex machina: Introduction to big data,” vol. 43, no. 1, pp. 19–39. [Online]. Available: https://doi.org/10.1146/annurev-soc-060116-053457

[2] D. Ruths, “The promises and pitfalls of demographic inference on social media.”

[3] P. Barber, “Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data,” vol. 23, no. 1, pp. 76–91. [Online]. Available: http://pan.oxfordjournals.org/content/23/1/76.abstract

[4] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, “Personality, gender, and age in the language of social media: The open-vocabulary approach,” vol. 8, no. 9, p. e73791. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0073791