Predicting Race and Ethnicity From Sequence of Characters in a Name[*]

Gaurav Sood[†]   Suriyan Laohaprapanon[‡]

April 22, 2018

Abstract

To answer questions about racial inequality, we often need a way to infer race and ethnicity from name. Until now, a bulk of the focus has been on optimally exploiting the last names list provided by the Census Bureau. But there is more information in the first names, especially for African Americans. To estimate the relationship between full names and race, we exploit the Florida voting registration data and the Wikipedia data (Ambekar et al. 2009). In particular, we model the relationship between sequence of characters in a name and race and ethnicity using Long Short Term Memory Networks. Our out of sample (OOS) precision and recall for the full name model estimated on the Florida Voter Registration data is .83 and .84 respectively. This compares to OOS precision and recall of .78 and .79 for the last name only model. Commensurate numbers for Wikipedia data are .72 and .73 for full name and .65 and .65 for last name models. To illustrate the use of this method, we apply our method to the campaign finance data to estimate the share of donations made by people of various racial groups.

1

How often are people of different races and ethnicites covered in the news? How often do African Americans contribute to political campaigns? To answer these questions and questions like these, we often need a way to infer race and ethnicity from names. Given the number of important questions at stake, numerous scholars have worked on inferring race from names. A bulk of the attention hitherto has been devoted to exploiting information in the last names list provided by the Census Bureau (see, for e.g., Fiscella and Fremont 2006; Imai and Khanna 2016). These efforts suffer from one major problem—lack of first names.

Lots of important uses cases, e.g., campaign donation records, voter registration records, carry both the first and the last name of a person. And we could exploit the information in the first name to make better predictions about the person's race and ethnicity. The information in the first name is especially vital for African Americans, whose last names are hard to distinguish from whites, and whose first names tend to distinctive (Bertrand and Mullainathan 2004).

In this paper, we exploit a novel source of data, the Florida Voter Registration Data for 2017, and Wikipedia data assembled by Ambekar et al. (2009), to build a model that estimates the relationship between sequence of characters in a name and race and ethnicity. We use Long Short Term Memory to learn the association between sequence of characters and race and ethnicity of a person. Our out of sample (OOS) precision and recall for the full name model estimated on the Florida registration data is .83 and .84 respectively. This compares to OOS precision and recall of .78 and .79 for the last name only model. Commensurate numbers for Wikipedia data are .72 and .73 for full name and .65 and .65 for last name models. We illustrate the use of our method by applying it to the campaign finance data to estimate the share of donations made by people of various racial groups. We also plan to investigate whether people are more likely to contribute to co-ethnics conditional on ideology, and descriptive information on racial composition of public employees. Lastly, we also provide a Python package to easily predict the race and ethnicity of names using

the models developed in this paper.

Data

We exploit two large datasets. Our first dataset is Florida Voting Registration data for the year 2017 (Sood 2017). The Florida Voting Registration for 2017 has information on nearly 13M voters along with their race. Given that we have very little data on voters who are multi-racial and who are Native Americans, we eliminate them from the data. Our final dataset only has information on voters who are Asian/Pacific Islander, Hispanic, Non-Hispanic Blacks, and Non-Hispanic Whites (see 1).

Table 1: Registered Voters in Florida by Race.

race	n
asian	253,808
hispanic	2,179,106
nh black	1,853,690
nh white	8,757,268

The Wikipedia data were originally collected by a team lead by Steven Skiena as part of the project to build a classifier for race and ethnicity based on names. The team scraped Wikipedia to produce a novel database of over 140k name–race associations (see 2). For details of the how the data was collected, see Ambekar et al. (2009). The dataset only contains unique names and can be seen as a sample of names of famous people. On the plus side, the Wikipedia data codes race at a much finer level—at a race, geographic region or religion level.

To derive some baselines, we also use the Census Bureau last name data. The Census Bureau provides frequency of all surnames occurring 100 or more times for the 2000 and 2010 census. Technical details of how the 2000 and 2010 data were collected can be found on the census website.

Table 2: Number of unique names by race in the Wikipedia Dataset.

race	n
Asian,GreaterEastAsian,EastAsian	5,497
Asian,GreaterEastAsian,Japanese	7,334
Asian,IndianSubContinent	7,861
GreaterAfrican,Africans	3,672
GreaterAfrican,Muslim	6,242
GreaterEuropean,British	41,445
GreaterEuropean,EastEuropean	8,329
GreaterEuropean,Jewish	10,239
GreaterEuropean,WestEuropean,French	12,293
GreaterEuropean,WestEuropean,Germanic	3,869
GreaterEuropean,WestEuropean,Hispanic	10,412
GreaterEuropean,WestEuropean,Italian	11,867
GreaterEuropean,WestEuropean,Nordic	4,813

We can use the Wikipedia data and the Florida Voting Registration data as is but the Census data needs to be transformed before being used. The dataset that the Census Bureau issues aggregates data for each last name and provides percentage of people with the last name who are Black, White, Asian, Hispanic, etc. Given some names are more common than others (Smith is the last name of 2,376,206 Americans), and given our interest in modeling the population distribution, we take a weighted random sample from this data with weight equal to how common the last name is in the population. Next, we assign race to name roughly in proportion to how the name is distributed across the racial groups. We assign floor of pctwhite as whites, floor of pctblacks as blacks etc. And we lose the one or two or few observations as we are using floor. We use this as the final dataset.

Model and Validation

For our full name model, we concatenate the last name and first name and capitalize the first character of all the words. We next split the name into two character chunks (bi-chars). For instance, Smith becomes Sm, mi, it, and th. We then remove infrequent bi-chars (occurring

less than 3 times in the data) and very frequent bi-chars (occurring more than 30% in the dataset). We use the remaining bi-chars as our vocabulary. We pad the sequences so that they are the same size. And we use 20 as the window size for the last name only model and 25 for the full name model. Lastly, we split the data randomly into train (80%) and test (20%). We next train a LSTM model.

The table 3 presents some metrics that shed light on how well we did with the last name only model in predicting race OOS. The table 4 presents some metrics that shed light on how well we did with the full name only model in predicting race OOS.

Table 3: Performance of the Last Name LSTM Model on the Florida Voter Registration Data.

race	precision	recall	f1-score	support
asian	0.77	0.35	0.48	50,762
hispanic	0.80	0.81	0.80	435,821
nh black	0.68	0.13	0.22	370,738
nh white	0.80	0.94	0.86	1,751,454
avg / total	0.78	0.79	0.75	2608775

Table 4: Performance of the Full Name LSTM Model on the Florida Voter Registration Data.

race	precision	recall	f1-score	support
asian	0.77	0.45	0.56	50,762
hispanic	0.83	0.84	0.83	435,821
nh black	0.74	0.45	0.56	370,738
nh white	0.86	0.93	0.89	1,751,454
avg / total	0.83	0.84	0.83	2608775

Given poor African Americans have distinctive first names, the largest improvements in recall that we see are among African American names. We go from a poor recall of .13 for African American names in the last name model to a still poor but much improved recall of .45 with the full name model.

Application

To illustrate how the package can be used, we impute the race of the campaign contributors recorded by FEC for the years 2000 and 2010 using the DIME data (Bonica 2017) and tally campaign contributions by race. In 2000, nearly 96% of the contributions were made by whites and just .16% of the contributions were made by blacks (see Table 5). In 2010, things didn't look a lot different, with nearly 94% of the contributions being made by whites and just .22% made by blacks.

Table 5: Proportion of Donors to Political Campaigns in 2000 and 2010 by Race.

race	2000	2010
asian	1.43%	2.14%
black	0.16%	.22%
hispanic	2.57%	4.04%
white	95.84%	93.60%

Discussion

We exploit a novel source of data—voter registration data—to learn a model between sequences of characters in a name and race or ethnicity. Given poor African Americans tend to have distinctive first names, the biggest advantage in using the full name model is in our ability to detect African American names. We use the model to infer the race of contributors in the DIME data and find that African Americans are less than a quarter percent of the donors. As we note, we also provide a Python package that exposes the models: https://github.com/appeler/ethnicolr/.

If you picked a random individual with last name Smith from the US in 2010 and asked us to guess this person's race (measured as crudely as by the census), the best guess would be based on what is available from the aggregated Census file. It is the Bayes Optimal Solution. So what good are last name only predictive models for? A few things. If you want

6

to impute ethnicity at a more granular level, guess the race of people in different years (than when the census was conducted if some assumptions hold), guess the race of people in different countries (again if some assumptions hold), when names are slightly different (again with some assumptions), etc. The big benefit comes from when both the first name and last name is known.

The limitations of using the voter registration data are obvious. Not everyone is registered to vote, and blacks and hispanics are especially likely to not be registered to vote (Ansolabehere and Hersh 2011). If the names of those who are not on the voter registration file are systematically different from those who are, we are likely somewhat optimistic in our accuracy metrics. Another concern with using data from a single state is that the pattern of names in a single state are different from names given in other states. It is a very reasonable concern. We could overcome it by combining census last name models with state voter registration data models but more research is needed to see how well we can do.

References

Ambekar, Anurag, Charles Ward, Jahangir Mohammed, Swapna Male and Steven Skiena. 2009. Name-ethnicity classification from open sources. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining.* ACM pp. 49–58.

Ansolabehere, Stephen and Eitan Hersh. 2011. "Gender, race, age, and voting: A research note.".

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American economic review* 94(4):991–1013.

Bonica, Adam. 2017. "Database on ideology, money in politics, and elections (DIME).".

Fiscella, Kevin and Allen M Fremont. 2006. "Use of geocoding and surname analysis to estimate race and ethnicity." *Health services research* 41(4p1):1482–1500.

Imai, Kosuke and Kabir Khanna. 2016. "Improving ecological inference by predicting individual ethnicity from voter registration records." *Political Analysis* 24(2):263–272.

Sood, Gaurav. 2017. "Florida Voter Registration Data.".
   URL: *https://doi.org/10.7910/DVN/UBIG3F*