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

Gaurav Sood[†] Suriyan Laohaprapanon[‡]

April 24, 2018

Abstract

To answer questions about racial inequality, we often need a way to infer race and ethnicity from a 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 voter registration data and the Wikipedia data (Ambekar et al. 2009). In particular, we model the relationship between the 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 the 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.

[*] Data and scripts behind the analysis presented here can be downloaded from http://github.com/appeler/ethnicolr.

[†] Gaurav can be reached at gsood07@gmail.com

[‡] Suriyan can be reached at: suriyant@gmail.com

How often are people of different races and ethnicities 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, 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 the sequence of characters in a name and race and ethnicity. We use Long Short Term Memory to learn the association between the 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 the 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 the 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 Table 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 Table 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 the 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 the 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 the floor of `pctwhite` as proportion white, floor of `pctblacks` as proportion black, etc. And we lose the one or two or few observations as we are using the floor. We use this as the final dataset.

Model and Validation

To learn the association between the sequence of characters in names and race and ethnicity, we estimate a LSTM model (Graves and Schmidhuber 2005; Gers, Schmidhuber and

Cummins 1999) on approximately 1M randomly sampled names from the Florida Voter Registration Data and about 100,000 randomly sampled name from the Wikipedia data. We estimate the last name model on an untransformed last name string. For the full name model, we concatenate the last name and first name but leave the capitalizations. We next split the string (last name or last name and first name) into two character chunks (bi-chars). For instance, Smith becomes `Sm, mi, it, th`. We then remove infrequent bi-chars (occurring less than 3 times in the data) and very frequent bi-chars (with 30% in the dataset). We use the remaining bi-chars as our vocabulary. In the Florida Voting Registration Data, this leaves us with 1,423 bi-chars in the case of last name only data, and 1,604 bi-chars in the full name data. In the Wikipedia data, the corresponding numbers are 1,946 and 2,260. Next, we pad the sequences so that they are the same size. Finally, we use 20 as the window size for the last name only model and 25 for the full name model.

On this set of sequences, we train a LSTM model using Keras (Chollet et al. 2015) and TensorFlow (Abadi et al. 2016). Before estimating the LSTM model, we embed each of the words onto a 32 length real-valued vector. We then estimate a LSTM with a .2 dropout (Srivastava et al. 2014) and .2 recurrent dropout for regularization. The last layer is a dense layer with a softmax activation which ensures that the probabilities of each of the classes sum of 1. Because it is a classification problem, we use log loss as the loss function and use ADAM for optimization. We fit the model for 15 epochs with a batch size of 32.

Table 3 presents some metrics that shed light on how well we did with the last name only model in predicting race OOS using the Florida Voter Registration Data. The OOS precision is .78, recall is .79, and f1-score is .75. There is however sizable variation in recall across different racial and ethnic groups. For instance, recall is .94 for whites and just .13 for non-Hispanic blacks.[1]

[1] You see the same pattern when we estimate the model on the Census last name data. Recall for blacks on the model estimated on both the 2000 and 2010 Census last name data is just .02 (see Table 9 and Table 10).

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

Compared to the last name only model, we do much better with a full name model. The OOS precision, recall, and f1-score for the full name model is .83, .84, and .83 respectively (see Table 4). The gains are, however, asymmetric. Recall is considerably better for Asians and Non-Hispanic blacks with the full name—.45 each, compared to .35 and .13 respectively. The gains in recall for Non-Hispanic blacks are expected given that African Americans have distinctive first names. The accuracy with which we predict non-Hispanic Blacks and Whites is also considerably higher—precision is 6 points higher for both non-Hispanic Blacks and Whites. Given Asians and Hispanics have more distinctive last names, the improvement in precision in predicting both is lower—negligible in case of Asians and 3 points in case of Hispanics.

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	260,8775

Moving to Wikipedia, the metrics look less pleasing than for the models based on the Florida voter registration data. This is expected. We have much less data and many more categories in the Wikipedia data. As Table 5 shows, the OOS precision, recall, and f1-score for the last name only model is .65 each respectively. For the full name model, the metrics are considerably better. The precision, recall, and f1-score jump to .72, .73, and .72 respectively (see Table 6).

Table 5: Performance of the Last Name LSTM Model on the Wikipedia Data.

race	precision	recall	f1-score	support
Asian,GreaterEastAsian,EastAsian	0.76	0.74	0.75	1,099
Asian,GreaterEastAsian,Japanese	0.82	0.84	0.83	1,467
Asian,IndianSubContinent	0.65	0.67	0.66	1,572
GreaterAfrican,Africans	0.50	0.36	0.42	734
GreaterAfrican,Muslim	0.56	0.52	0.54	1,248
GreaterEuropean,British	0.72	0.85	0.78	8,289
GreaterEuropean,EastEuropean	0.75	0.62	0.68	1,666
GreaterEuropean,Jewish	0.43	0.37	0.40	2,048
GreaterEuropean,WestEuropean,French	0.56	0.47	0.51	2,459
GreaterEuropean,WestEuropean,Germanic	0.38	0.29	0.33	774
GreaterEuropean,WestEuropean,Hispanic	0.63	0.49	0.55	2,082
GreaterEuropean,WestEuropean,Italian	0.61	0.74	0.67	2,374
GreaterEuropean,WestEuropean,Nordic	0.65	0.53	0.58	963
avg / total	0.65	0.65	0.65	26,775

Table 6: Performance of the Full Name LSTM Model on the Wikipedia Data.

race	precision	recall	f1-score	support
Asian,GreaterEastAsian,EastAsian	0.86	0.80	0.83	1,099
Asian,GreaterEastAsian,Japanese	0.89	0.90	0.90	1,467
Asian,IndianSubContinent	0.78	0.75	0.76	1,572
GreaterAfrican,Africans	0.56	0.41	0.47	734
GreaterAfrican,Muslim	0.64	0.67	0.66	1,248
GreaterEuropean,British 0.76	0.88	0.81	8,289	
GreaterEuropean,EastEuropean	0.76	0.74	0.75	1,666
GreaterEuropean,Jewish	0.51	0.42	0.46	2,048
GreaterEuropean,WestEuropean,French	0.69	0.61	0.65	2,459
GreaterEuropean,WestEuropean,Germanic	0.49	0.43	0.46	774
GreaterEuropean,WestEuropean,Hispanic	0.73	0.69	0.71	2,082
GreaterEuropean,WestEuropean,Italian	0.75	0.74	0.75	2,374
GreaterEuropean,WestEuropean,Nordic	0.75	0.62	0.68	963
avg / total	0.72	0.73	0.72	26,775

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 7). 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 7: Proportion of Donations 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%

It may be that blacks make much fewer number of donations as compared to whites but may donate a larger amount when they donate. What proportion of the total donated sum comes from people from different racial groups? As Table 8 shows, the numbers are virtually unchanged.

Table 8: Proportion of Total Amount Donated to Political Campaigns in 2000 and 2010 by Race.

race	2000	2010
asian	1.55%	2.11%
black	0.16%	.17%
hispanic	2.49%	2.82%
white	95.81%	94.90%

Discussion

We exploit a novel source of labeled data—voter registration files—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 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. And there are a lot of important datasets, such as the campaign contributions dataset, the voter registration files of other states, news data, etc., where we have information on both the first and the last names. And we could make better predictions about the race and ethnicity by capitalizing on both the first and the last names, especially for African Americans, but also other ethnicities.

The limitations of using the voter registration data are obvious. Not everyone is registered to vote, and blacks and Hispanics are especially likely not to 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

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard et al. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *OSDI*. Vol. 16 pp. 265–283.

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).".

Chollet, François et al. 2015. "Keras.".

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.

Gers, Felix A, Jürgen Schmidhuber and Fred Cummins. 1999. "Learning to forget: Continual prediction with LSTM.".

Graves, Alex and Jürgen Schmidhuber. 2005. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures." *Neural Networks* 18(5-6):602–610.

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*

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov. 2014. “Dropout: A simple way to prevent neural networks from overfitting.” *The Journal of Machine Learning Research* 15(1):1929–1958.

Appendix: Census Models

Table 9: Performance of the Last Name LSTM Model on the Census 2000 Data.

race	precision	recall	f1-score	support
api	0.84	0.64	0.73	6,994
black	0.63 0.02	0.04	25,176	
hispanic	0.86	0.83	0.85	25,629
white	0.82	0.98	0.89	142,201
avg / total	0.80	0.83	0.77	200,000

Table 10: Performance of the Full Name LSTM Model on the Census 2010 Data.

race	precision	recall	f1-score	support
api	0.82	0.56	0.66	10,114
black	0.64	0.02	0.04	24,807
hispanic	0.87	0.76	0.81	32,923
white	0.78	0.97	0.87	132,156
avg / total	0.78	0.80	0.74	200,000