`demographicx`: A Python package for estimating gender and ethnicity using deep learning transformers

Lizhen Liang and Daniel E. Acuna

School of Information Studies, Syracuse University, Syracuse, NY

Abstract

Plenty of research questions would benefit from understanding whether demographic factors are associated with social phenomena. Accessing this information from individuals is many times infeasible or unethical. While software packages have been developed for inferring this information, they are often untested, outdated, or with licensing restrictions. Here, we present a Python package to infer the gender and ethnicity of individuals using first names or full names. We employ a deep learning transformer of text fragments based on BERT to fine-tune a network. We train our model on Torkiv (Torvik 2018), and extensively validate our predictions. Our gender prediction achieves an average F1 of 0.942 across female, male, and unknown gender names. Similarly, our ethnicity prediction achieves an average F1 of 0.94 across White, Black, Hispanic, and Asian categories. We hope that by making our package open and tested, we improve demographic estimates for many research fields that are trying to understand these factors.

Statement of Need

Demographic information such as gender and ethnicity is a crucial dimension to understand many social phenomena. Gender and ethnicity are of course only a fraction of the critical factors that should be analyzed about individuals (see (Acuna 2020)), yet they have attracted increased interest from the research community. In social science, for example, it has been shown that gender and race are important for scientific collaboration (Larivière et al. 2013), mentorship (Schwartz, Liénard, and David 2021), and funding (Ginther et al. 2011). Accessing this information is, however, challenging because of legal or ethical reasons. Many studies resort to analyzing names to make these kinds of inferences, but the packages and services they often use are non-reproducible or rely on proprietary information with unknown methods and validations (e.g.,

genderize.io). Without access to an easy-to-use, public, open, and validated method, we risk making inferences about these kinds of phenomena without good grounding. While inferring demographics from names has potential flaws (Kozlowski et al. 2021), it is sometimes the only input we have; it is desirable to have better algorithms than the ones currently available.

Table 1: Gender prediction performance on validation split of the mixed data set and Social Security Administration (SSA) popular newborn names. Names in SSA that are also in validation and with a "unknown" label in authori-ty data set will take the label from authori-ty in order to validate performance on "unknown" class.

	Male		**Female**		**Unknown**	
	Validation	SSA	Validation	SSA	Validation	SSA
F1	0.961	0.813	0.975	0.915	0.889	0.504
Acc	0.972	0.711	0.979	0.885	0.862	0.664
AUC	0.993	0.954	0.996	0.965	0.966	0.860

Here, we describe a Python package called `demographicx` which infers gender from first name and ethnicity from the full name. It is based on fine-tuning a deep learning BERT embedding model with sub-word tokenization (Devlin et al., 2018). Importantly, our model has the ability to make predictions for names that it has not seen before. We build our package on top of the popular transformers package, which increases the likelihood that users will have parts of our models cached in their computers. The dataset we used to train includes Genni + Ethnea for the Author-ity 2009 dataset by Torvik (Torvik 2018), which has names and predicted results by other previous methods. We mixed the dataset with the Social Security Administration (SSA) popular newborn baby names dataset (Social Security Administration 2013) and a Wikipedia name ethnicity dataset (Ambekar et al. 2009). We validate our model with both the aggregated data set and the Wikipedia datasets. Our models achieve excellent performance on both tasks (see Table 1 and 2).

Table 2: Race prediction performance on validation (val) split of the mixed data set and Wikipedia (Wiki) names

	Black		Hispanic		White		Asian	
	Val	Wiki	Val	Wiki	Val	Wiki	Val	Wiki
F1	0.976	0.987	0.936	0.822	0.907	0.850	0.941	0.859
Acc	0.999	0.999	0.928	0.788	0.902	0.856	0.931	0.843
AUC	0.999	0.996	0.990	0.964	0.983	0.963	0.989	0.962

Because our package is built based on the `transformers` package, it can be

easily incorporated into PyTorch and transformers. The API is very simple on purpose. Our package has already been used in (Acuna and Liang 2021) and multiple other internal projects.

```
In: from demographicx import GenderEstimator
In: gender_estimator = GenderEstimator()
In: gender_estimator.predict("Daniel")
Out: {'female': 0.001, 'male': 0.988, 'unknown', 0.011}

In: gender_estimator.predict("Amy")
Out: {'female': 0.998, 'male': 0.001, 'unknown', 0.001}

In: from demographicx import EthnicityEstimator
In: ethnicity_estimator = EthnicityEstimator()
In: ethnicity_estimator.predict("Daniel Acuna")
Out: {'white': 0.002, 'hispanic': 0.998, 'black', 0.000, 'asian': 0.000}

In: ethnicity_estimator.predict("Lizhen Liang")
Out: {'white': 0.000, 'hispanic': 0.000, 'black', 0.000, 'asian': 0.999}
```

Acknowledgments

L. Liang and D. Acuna were partially funded by the National Science Foundation grant "Social Dynamics of Knowledge Transfer Through Scientific Mentorship and Publication" #1933803. We thank Jim Yi for his help with the repository.

References

Acuna, Daniel E. 2020. "Some Considerations for Studying Gender, Mentorship, and Scientific Impact: Commentary on Alshebli, Makovi, and Rahwan (2020)." *OSF Preprints*. https://doi.org/10.31219/osf.io/ybfk6.

Acuna, Daniel E, and Lizhen Liang. 2021. "Are Ai Ethics Conferences Different and More Diverse Compared to Traditional Computer Science Conferences?" In *Fourth Aaai/Acm Conference on Artificial Intelligence, Ethics, and Society (Aies'21)*. Association for Computing Machinery. https://doi.org/10.1145/3461702.3462616.

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*, 49–58. https://doi.org/10.1145/1557019.1557032.

Ginther, Donna K, Walter T Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L Haak, and Raynard Kington. 2011. "Race, Ethnicity, and Nih Research Awards." *Science* 333 (6045). American Association for the Advancement of

Science: 1015–9. https://doi.org/10.1126/science.1196783.

Kozlowski, Diego, Dakota S Murray, Alexis Bell, Will Hulsey, Vincent Larivière, Thema Monroe-White, and Cassidy R Sugimoto. 2021. "Avoiding Bias When Inferring Race Using Name-Based Approaches." *arXiv Preprint arXiv:2104.12553*.

Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R Sugimoto. 2013. "Bibliometrics: Global Gender Disparities in Science." *Nature News* 504 (7479): 211. https://doi.org/10.1038/504211a.

Schwartz, Leah P, Jean Liénard, and Stephen V David. 2021. "Impact of Gender on the Formation and Outcome of Mentoring Relationships in Academic Research." *arXiv Preprint arXiv:2104.07780*.

Social Security Administration. 2013. "Beyond the Top 1000 Names." *Retrieved March* 20: 2014.

Torvik, Vetle. 2018. "Genni + Ethnea for the Author-Ity 2009 Dataset." University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-9087546_V1.