A Name-Based Race Classifier Using Collaborations in Academia

Mathilde Simoni

Department of Conputer Science, New York University Abu Dhabi

Introduction

Race is an important categorization standing as a proxy to represent a wide range of cultures, languages, experiences, and a factor often considered in biomedical, sociological, and ethnographic research. However, gathering data about race is challenging due to privacy concerns. Current methods use name-based classifiers, machine learning techniques to predict race from names. Nevertheless, they have limited availability and accuracy for under-represented communities.

To fill this gap, we propose a novel approach for race classification, using name embedding and collaborations in academia. It has been shown that homophily patterns, the tendency to communicate with similar others, exist in academia and are especially strong with respect to race and ethnicity [1]. We use collaborations to encode the racial meaning of names in the form of name embedding.

"Race is a factor often considered in biomedical, sociological, and ethnographic research"

Methodology

1. **Generation of name embedding**: we use the MAG (Microsoft Academic Graph) dataset to create a graph of collaborations and later generate name embedding for a large diversity of first and last names.
2. **Creation of a dataset with full names**: we generate a list of synthetic full names for each race by combining first and last names from a US voters' dataset.
3. **Classification**: we use a K-Nearest-Neighbors algorithm to classify full names into 4 race categories (defined by the US Census Bureau as White, Black, Hispanic, or Asian).

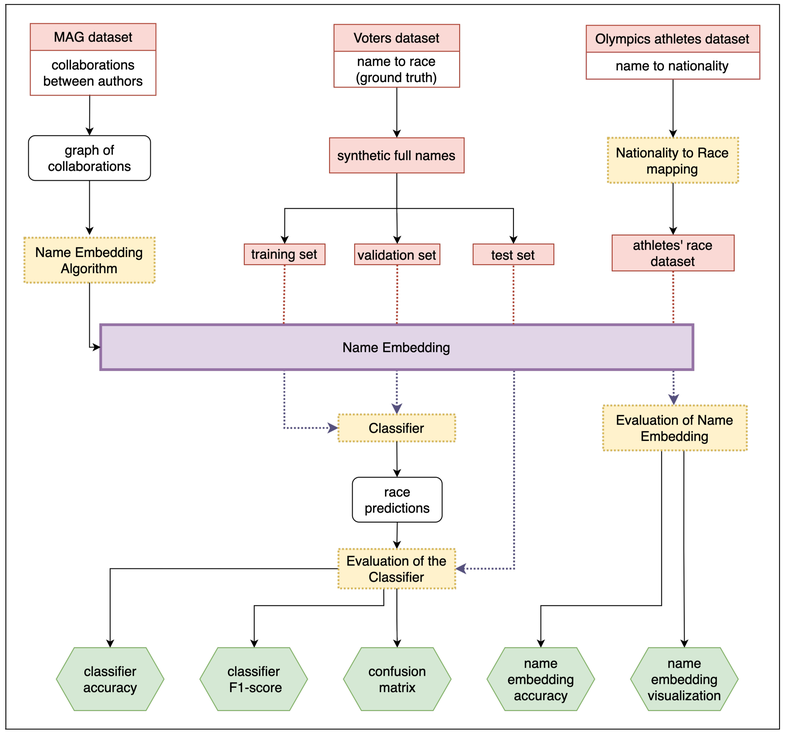

Figure 1. Methodology

__Name embedding__ is a form of word embedding, an abstract representation in an n-dimensional space where words that co-occur frequently in their context are represented by points close to each other in space. In this context, names represented by adjacent points are highly likely to belong to the same race group.

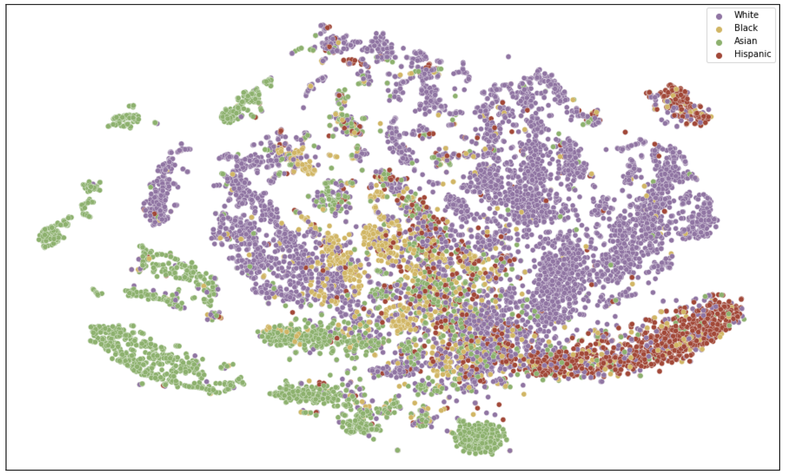

Figure 2. First name embedding visualization

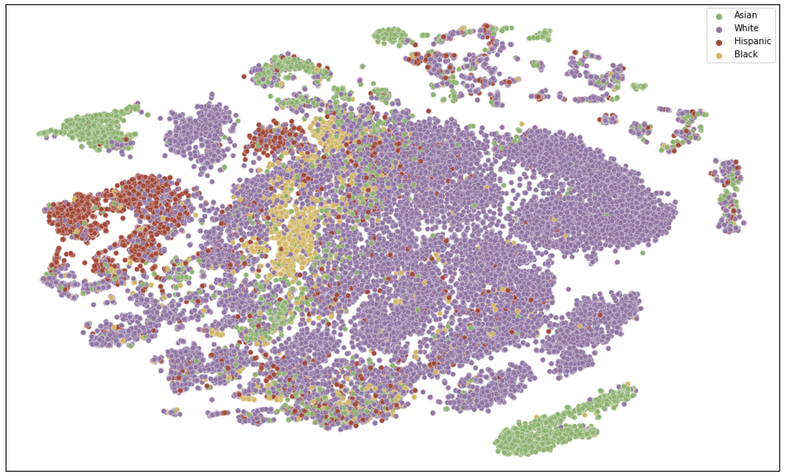

Figure 3. Last name embedding visualization

Evaluation of Name Embedding

We evaluate the accuracy of the name embedding with a dataset of Olympics athletes. We infer race from the country of representation, as athletes usually represent their country of origin. The embedding is **70% accurate for first names, 77% accurate for last names** and assigns a representation to the same proportion of White, Black, Asian, and Hispanic names.

Figures 2 and 3 display visualizations of the name embedding for all first and last names in the Olympics dataset. We use PCA [2] and t-SNE [4] methods to project the 100D name embeddings into 2D. Distinct clusters can be observed for each race category, with more noticeable regions for Asian and Hispanic names.

Results

Accuracy: 95.45%
F1-score: 0.955

We compare these results with Namsor, currently ranked as the world's most accurate name verification technology. On a benchmark set on 1000 names, we obtain an accuracy of 95.70% (F1-score of 0.957), while Namsor reaches an accuracy of 76.90% and an F1-score of 0.766. Therefore, **we perform better than one of the best name-based classifiers currently on the market.**

In addition, table 1 shows that the predictions are fair across races, as the F1-score and accuracy obtained are very similar for each race category.

Race	F1-score	Accuracy
White	0.952	94.8%
Black	0.942	91.3%
Hispanic	0.986	97.5%
Asian	0.957	98.3%

Table 1. Accuracy and F1-score for each race group

Discussion

How were fairness and representation addressed in this project?

- We generated an embedding that approximately covers the same percentage of White, Black, Hispanic, and Asian names.
- We used balanced training, validation, and test sets.
- We achieved predictions with very similar accuracy and F1-score for each race group.

What is this study's main takeaway?

Using academic collaborations to learn name embedding is an effective way to capture the underlying racial dimension of names, confirming the strong presence of homophily patterns with regard to race in the academic sphere.

Acknowledgements

I would like to express my deepest gratitude to Fengyuan "Michael" Liu for invaluable and constant support throughout the 3 semesters, whether it be for technical issues or general help for the project. I also thank my advisors Talal Rahwan and Bedoor AlShebli for their guidance and encouragement.

References

[1] Bedoor K AlShebli, Talal Rahwan, and Wei Lee Woon.
The preeminence of ethnic diversity in scientific collaboration.
Nature communications, 9(1):5163, 2018.

[2] Harold Hotelling.
Analysis of a complex of statistical variables into principal components.
Journal of educational psychology, 24(6):417, 1933.

[3] Evan TR Rosenman, Santiago Olivella, and Kosuke Imai.
Race and ethnicity data for first, middle, and last names.
arXiv preprint arXiv:2208.12443, 2022.

[4] Laurens Van der Maaten and Geoffrey Hinton.
Visualizing data using t-sne.
Journal of machine learning research, 9(11), 2008.

mps565@nyu.edu