A Name-Based Race Classifier Using Name Embedding and Collaborations in Academia

Mathilde Simoni

Capstone Project, May 2023

Outline

- Literature
- Methodology
  - Creation of the Graph of Collaborations
  - Name Embedding Algorithm
  - Generation of Synthetic Full Names
  - The classifier
- Evaluation of Name Embedding
- Results
- Discussion
- Conclusion

Literature

Nameprism

- Relies on the principle of homophily in communication patterns
- Uses contact lists from a major internet company to learn name embedding
- Naive-Bayes approach
- F1-score of 0.795

Weakness: lower accuracy for Black names

Namsor

- Most accurate name verification technology in the world
- Training set of more than 8 billion names
- Taxonomy of 142 ethnicities

Weakness: not available / free to use

The need:

A free classifier with a method focused on increasing the accuracy for underrepresented communities

PART 01

Methodology

Main idea: use the *homophily principle*, assuming that authors collaborate more often with other scholars of the same race

Creation of the Graph of Collaborations

- MAG dataset
- Graph
  - Node: author
  - Edge: collaboration
  - Weight: number of collaborations
- First names
  - 90,307,136 edges
  - 3,212,598 nodes
- Last names
  - 253,679,354 edges
  - 5,038,383 nodes

Name Embedding Algorithm

- Adjacent points are highly likely to belong to the same race group
- GRAPE software
- Node2Vec algorithm
- Embedding dimension: 100
- 30 epochs

Generation of Synthetic Full Names with the Voters' dataset

Method 1

1. Select first and last names assigned to a race with probability > 90%
2. Create all full name combinations for White, Black, Hispanic and Asian names
3. Select random samples (25% for each race)
4. **Split in training set (80%), test and validation sets (10% each)**

Method 2

1. Select first and last names assigned to a race with probability > 90%
2. **Split in training set (80%), test and validation sets (10% each)**
3. Generate full name combinations for training, test and validation sets separately
4. Select random samples (25% for each race)

The classifier

K-nearest-neighbors (40 neighbors for method 1 and 150 neighbors for method 2)

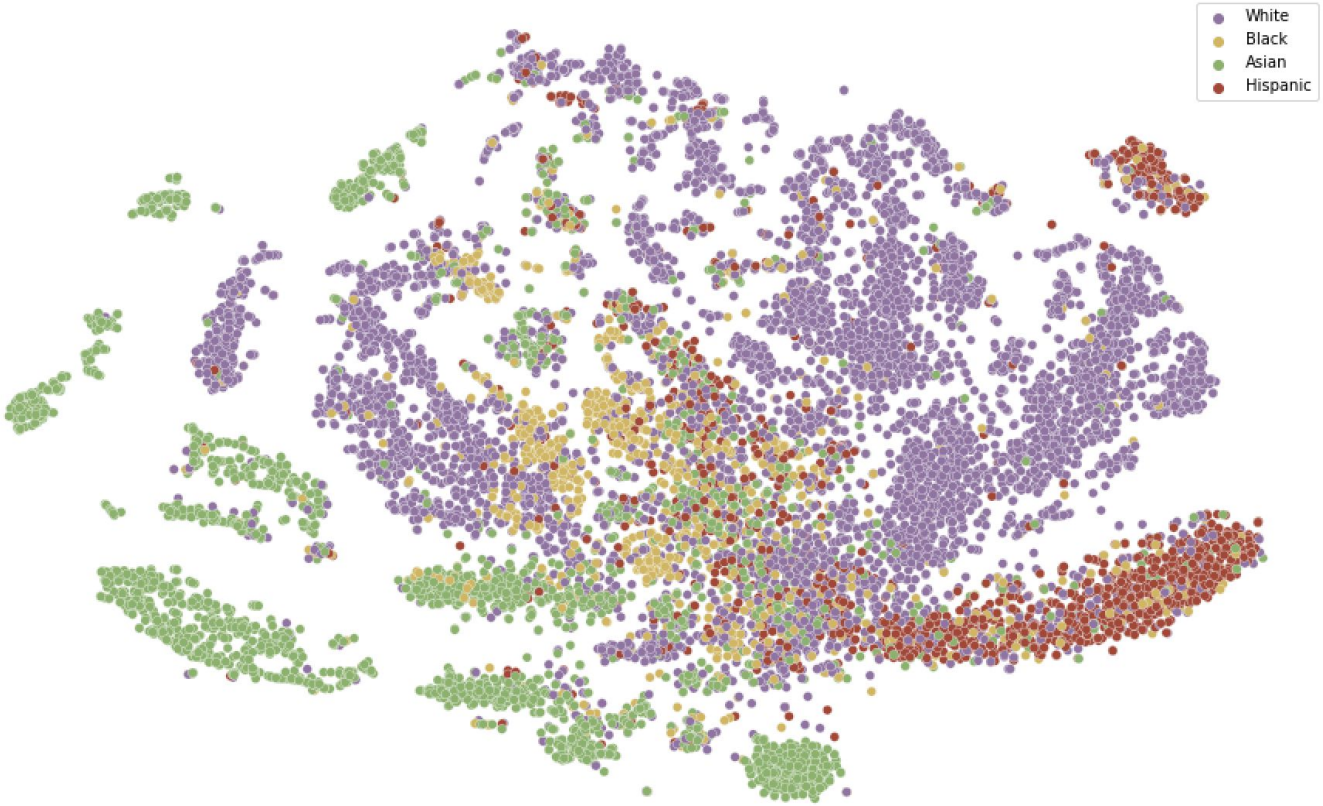
PART 02

Evaluation of Name Embedding

With the Olympics dataset

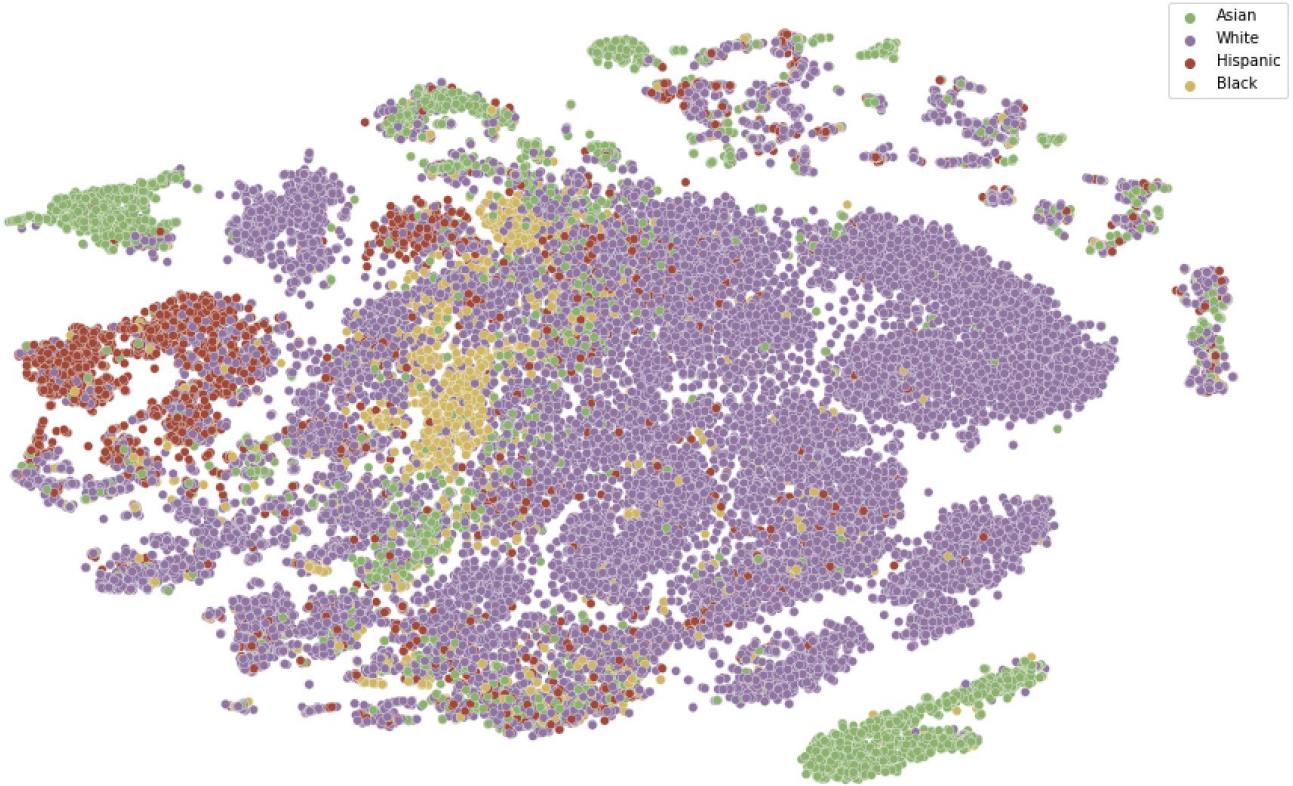
Race Representation

Race	% of first names	% of last names
White	90.7%	72.7%
Black	86.7%	76.6%
Asian	87.2%	63.9%
Hispanic	91.0%	89.2%



Embedding Visualization

First Names



Embedding Visualization

Last Names

Quantitative Study

- 10 nearest neighbors
- Accuracy:
  - 70% for first names
  - 77% for last names

Qualitative Study

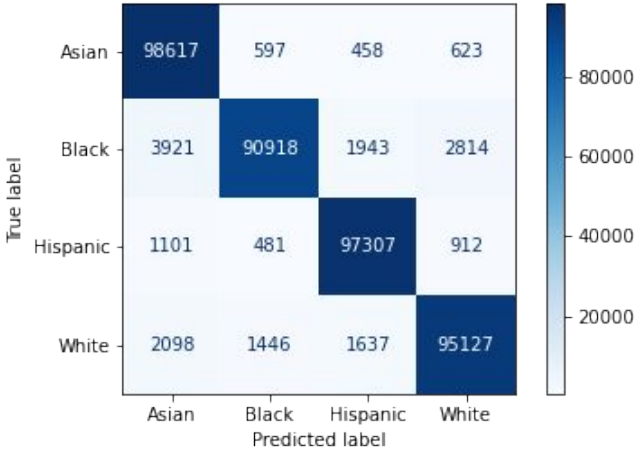
first name	10 nearest neighbors
mariam	samer, tariq, maha, marwa, amira, saad, ashraf, amr, habib, abdallah
haoyu	xinyuan, wenying, jihua, duo, xuebin, zili, jiajie, yuxuan, xiaoxi, yunhui
haruto	tsumoru, ikuji, toshinao, takesumi, ryonosuke, kimiyuki, kaneharu, mutsuki, kakuhiro, atsuto
jose	paula, fernando, eduardo, francisco, ricardo, diana, jorge, rafael, joao, diego
pierre	michel, ian, bruno, jean, francois, christophe, isabelle, bernard, gregory, olivier

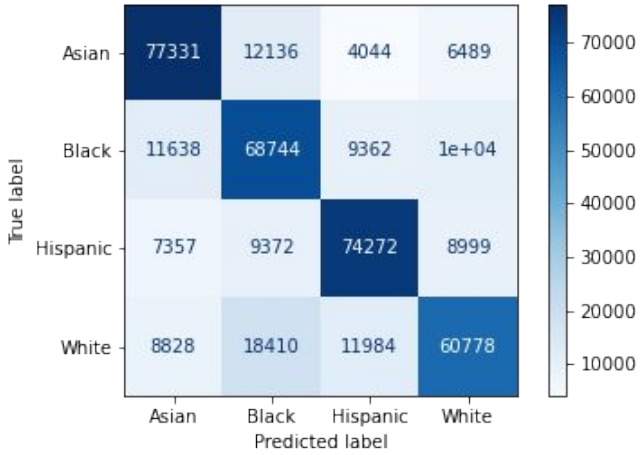
PART 03

Results

Method 1

Set	Our Classifier		Namsor	
	accuracy	F1-score	accuracy	F1-score
test	**95.49%**	**0.955**	/	/
benchmark	95.70%	0.957	76.90%	0.766





Method 2

Set	Our Classifier		Namsor	
	accuracy	F1-score	accuracy	F1-score
test	70.28%	0.721	/	/
benchmark	72.50%	0.723	**77.80%**	**0.770**

Discussion

- Method 1 *VS* Method 2
  - The distribution of values in a test set is supposed to be as close to the real world distribution as possible
  - Lower accuracy for method 2 highlights bias in method 1
  - Results are not too far from Namsor results

- More variety for White names

- Namsor uses country of residence to predict race

Conclusion

Using collaborations in academia to learn name embedding is an effective way to capture the underlying racial dimension of names. It confirms the strong presence of homophily patterns with regard to race in academia

Problem of representation

- Name embedding roughly covers the same proportion of White, Black, Hispanic and Asian names
- Balanced training set: same number of full names for each race group
- Balanced predictions among race categories

Next steps

- Include more names in the name to embedding mapping (closest distance algorithms)
- Use language models to improve name embedding
- Try other classifiers