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

Mathilde Simoni
Computer Science, NYUAD
mps565@nyu.edu

Advised by: Talal Rahwan, Bedoor AlShebli, Fengyuan "Michael" Liu

ABSTRACT

Names tell a lot about someone. They are indicators of gender, country of origin, religion, age, or parent's main fields of interest. In addition, they are perceived as the most practical way to gather race information. Today, inferring race from name is often needed for economic and political science research, but existing name-based race classifiers have limited accuracy and availability. They are often trained on unrepresentative datasets and are less accurate for underrepresented communities, or are not freely available.

To fill this gap, we propose a novel approach for race classification, using name embedding and collaborations in academia. Exploiting the phenomenon of homophily in communication patterns, we assumed that authors collaborate more frequently with other scholars of the same race. This allowed us to use the Microsoft Academic Graph, a dataset gathering data on collaborations of millions of authors, to learn name embedding and encode a wide variety of names. We subsequently generated lists of synthetic full names for each race category, combining first and last names from a US voters' dataset with self-reported race (as defined by the US Census Bureau). A K-Nearest-Neighbors classifier was trained and tested and obtained an accuracy of 95.45% with an F1-score of 0.955. It performed better than Namsor (accuracy of 76.90%), currently the most accurate name verification technology, on a benchmark set of a thousand names. In addition, the prediction errors were balanced across race groups. We believe that this tool will be useful for future research where race information is needed, but not available.

This report is submitted to NYUAD's capstone repository in fulfillment of NYUAD's Computer Science major graduation requirements.

جامعة نيويورك أبوظبي
NYU ABU DHABI

Capstone Project, Spring 2023, Abu Dhabi, UAE

KEYWORDS

race classification, name embedding, machine learning, academic collaborations, K-nearest-neighbors

Reference Format:
Mathilde Simoni. 2023. A Name-Based Race Classifier Using Name Embedding and Collaborations in Academia. In *NYUAD Capstone Project Reports, Spring 2023, Abu Dhabi, UAE*. 11 pages.

1 INTRODUCTION

Ethnicity and race are important categorizations, standing as a proxy to represent a wide range of cultures, languages, and experiences around the world. They are also factors often considered in biomedical, sociological, and ethnographic research. For instance, they are used to analyze genetic differences or public health disparities among different groups of people [1, 2] and to study discrimination [3]. The classification of ethnic groups has therefore many applications in academia. However, race data is usually not readily available and privacy concerns make the process of gathering data challenging in the field.

In the current methods, names are often perceived as the most practical way to gather race data other than direct questions. However, existing name-based race classifiers show limited performance. The earlier methods are trained on small and unrepresentative sets of labeled names or use substrings as features. Recently, *NamePrism* [4] and *Namsor* showed great improvement. However, *NamePrism* does not perform well for Black communities, due to the under-representation of Black individuals in the training data. Namsor is currently ranked as the more accurate name verification algorithm in the world to predict race, gender, country of origin, and ethnicity. Nonetheless, the methods used are not publicly available to the research community and the classifier is not free to use, including for academic purposes.

Therefore, there is a concrete need to develop free classifiers and methods focused on increasing the number of correct predictions for underrepresented communities. In

this paper, we contribute to the existing body of work by proposing a novel approach for race classification, using name embedding and collaborations in academia.

It has been shown that homophily patterns exist in academia and are especially strong with respect to race and ethnicity [5]. In other words, authors tend to collaborate more with people of the same race in order to publish academic papers. We use the MAG (Microsoft Academic Graph) dataset, which gathers data from millions of authors worldwide, to obtain data on collaborations among a great diversity of authors from different locations and race groups. Moreover, name embedding has already been used by the authors of *NamePrism* and has been proven to be very powerful to "encode homophily as features without explicit labels of gender, ethnicity and nationality" [6]. Therefore, using collaborations to create name embedding is a powerful means to encode the underlying racial aspect of names. We use this name embedding and a ground truth dataset of US voters to build a classifier that predicts race (defined by the US Census Bureau as White, Black, Hispanic, or Asian) based on full name.

2 RELATED WORK

The classification of individuals' race or ethnic group based on their names has been a topic of interest for researchers in various fields, including biomedical research, sociology, and natural language processing. Numerous approaches have been presented, and the most recent ones are described below.

Ethnea [7] uses an instance-based approach to name-based ethnicity classification, with 26 predefined ethnicity classes. The classifier compares the name and location of an individual to those of previously classified authors in a large-scale bibliographic database, comprising more than tens of millions of names distributed across 200+ countries and over 20+ years.

NamePrism [4] relies on the principle of homophily in communications, assuming that individuals communicate more frequently with other individuals of similar age, language, and location. Instead of using traditional substrings methods to encode the meaning of names, the classifier exploits contact lists from a major internet company to learn name embedding, an abstract representation where names communicating more often are represented close to each other in the embedding space. A naïve-based approach is subsequently used to classify the nationality of unlabeled names, which are then assigned an ethnicity using a taxonomy based on Cultural, Ethnic, and Linguist (CEL) similarities. The results show that *NamePrism* achieves high accuracy in predicting nationality, with an F1 score of 0.795. However, the classifier is weaker for Black names, as they represent only 3 % of the names in the contact lists used to generate name embedding.

Similarly, Junting Ye and Steven Skiena developed an ethnicity classifier using name embedding [6]. The main difference with *NamePrism* is that the training dataset was obtained from a publicly accessible Twitter database, under the assumption that homophily patterns in communications are universal and also exist in social media. The classifier was trained with 68 million Tweets and 89 million unique user profiles and obtained an accuracy of 73%. Furthermore, their study [6] proves the power of name embedding by showing that it can help improve demographic models, such as lifespan modeling, when other demographic features, including age or gender, are also available.

More recently, *rethnicity* [8], a free and user-friendly R-package used to predict ethnicity based on names, was developed. It uses bi-LSTM (Bidirectional Long-Short Term Memory) recurrent neural networks with a dataset obtained from Florida voters registration. In addition, it gives special care to the accuracy of predictions for minority groups, under-sampling the majority classes in the training dataset, hence adjusting its imbalances.

Currently, *Namsor* is ranked by science Metrix as the most accurate name verification technology in the world for a large diversity of international names, being able to correctly classify personal names by gender, country or origin, race and ethnicity. It uses a training dataset of more than 8 billion names, supports 22 alphabets and a complete taxonomy of 22 regions, 249 countries, 3 genders and 142 ethnicities. It was used to infer gender and ethnicity for a wide range of studies, including works on gender imbalances in academia [9–11] and other studies on diversity [12, 13].

3 DATASETS

3.1 MAG Dataset

The MAG (Microsoft Academic Graph) dataset is available at https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/. It was published by Microsoft and contains scientific publication records including information about authors, institutions, papers, as well journals, fields of study, and citations. In particular, the database gathers 86,965,155 unique full names, 9,257,240 unique first names, 13,028,415 unique last names, and data about 270,694,050 academic papers from 27,076 institutions. In the study, we used the file Authors.txt, containing unique Author IDs and their full names in normalized form, as well as PaperAuthorAffiliation.txt comprised of Author IDs and their published paper IDs. This data allowed us to get information about collaborations between authors in academia.

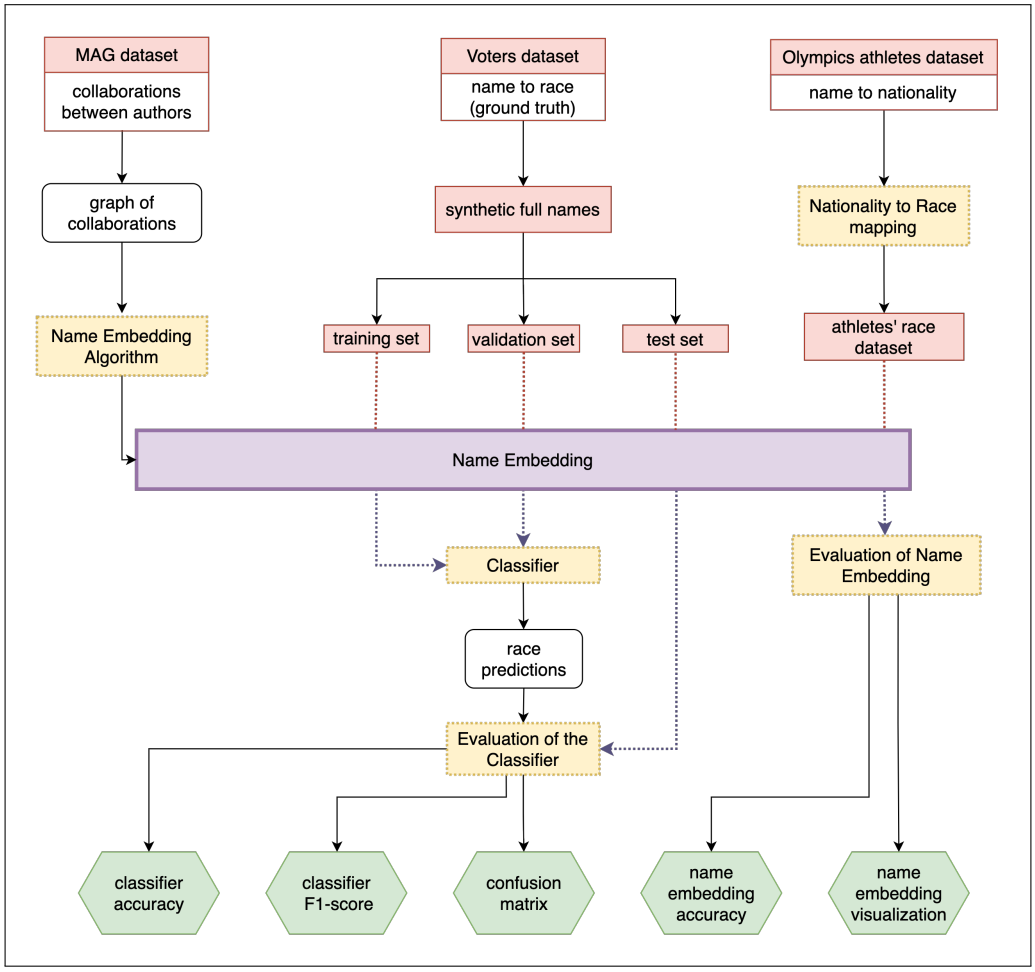

Figure 1: Creation and Evaluation of the Name-Based Race Classifier

3.2 Olympics Athletes Dataset

The Olympics Athletes dataset was extracted from Kaggle and is available at https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results. It gathers 134,723 unique full names of Olympic athletes who participated in the Games from Athens 1896 to Rio 2016. It provides each athlete's country of representation, a valuable attribute as Olympics participants usually represent their country of origin. Moreover, this dataset contains a wide variety of names from different cultures and ethnicities, the Olympics being a highly international event. After inferring the nationality of each athlete from their country of representation, we used this dataset to produce visualizations and calculate the accuracy of the name embedding results.

3.3 Voters' Dataset

This dataset [14] provides ground-truth race data collected from the voter files of 5 US states and was used to train and evaluate the classifier. It contains up to 1 million first, middle, and last names and self-reported race, classified into 5 race groups: White, Black, Hispanic, Asian, and Other. As mentioned in the original paper [14], this dataset can be used to obtain conditional probabilities of race given names or improve the accuracy of existing predictions. However, the authors also discuss potential biases as the names are not drawn from a random sample of the US population. First, voters are above 18 years old and can differ from non-voters in terms of economic status. In addition, the southern states of the US used to create this dataset have smaller Hispanic and Asian populations than the country average, potentially leading to bias against these communities in classifiers using this dataset as training data.

4 METHODOLOGY

Figure 1 shows a summary of the steps taken to create and evaluate the name-based race classifier.

4.1 Creation of the Graph of Collaborations

Academic collaborations can be studied by analyzing co-authorship. In MAG dataset, the two files "Authors.txt" and "PaperAuthorAffiliations.txt" gathering data about papers and their co-authors were used to create a graph of collaborations. A full name was represented by a node in the graph, and a collaboration between two authors named A and B (the act of publishing an academic paper as collaborators) was expressed by creating an edge between the nodes with names A and B. In addition, edges were assigned a weight to symbolize the number of collaborations between authors with the two full names. The resulting graph contained 1,420,152,952 edges.

Furthermore, the original graph was split into two graphs for first and last names. This separation was performed in order to later generate name embeddings for first and last names separately, hence account for a larger variety of full names in the training and test sets. In addition, names consisting of one character and names of authors collaborating only once were removed to reduce noise. This procedure resulted in two graphs with the following characteristics:

First names graph:

- Number of edges: 90,307,136
- Number of nodes: 3,212,598
- Maximum weight: 343,656
- Minimum weight: 2
- Average weight: 16

Last names graph:

- Number of edges: 253,679,354
- Number of nodes: 5,038,383
- Maximum weight: 3,478,302
- Minimum weight: 2
- Average weight: 44

4.2 Name Embedding Algorithm

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. Such representations have been extensively used in the field of natural language processing. For this study, we used the previously created graph of collaborations as input to the name embedding algorithm. Therefore, names associated with authors who frequently collaborate together are represented by adjacent points in embedding space.

This method was motivated by the phenomenon of homophily in communication patterns. The homophily principle describes how similar individuals, with regard to gender, race, religion, occupation, age, or other social identities, tend to communicate more often with each other. It structures network ties and influences how people build connections, leading toward more homogeneous groups. Among all factors, homophily regarding race and ethnicity "creates the strongest divides in [...] personal environments" [15]. In addition, it has been shown that homophily patterns can be observed in the academic sphere including at academic conferences [16] or in the context of co-authorship [5]. In particular, AlShebli et al. Found that "ethnic homophily is steadily increasing" [5] over time in academia.

Exploiting the phenomenon of homophily, we assumed that authors collaborate more frequently with other scholars of the same race. This allowed us to create a name embedding where names represented by adjacent points are highly likely to belong to the same race group. In other words, the locality property of this representation reflected an underlying similitude between names of the same race.

Regarding the implementation, we used GRAPE [17], a recent software for graph processing and representation learning designed to scale with big graphs (millions of nodes and billions of edges) and to handle parallel computing. It has been proven to have better space and time performance compared to other software, including graph libraries such as NetworkX, or algorithms such as Deep Walk, and its documentation can be found at https://anacletolab.github.io/grape/grape. GRAPE provides an optimized version of Node2Vec [18], an algorithm that computes the name embedding of nodes belonging to a graph. Its performance is better than the traditional Node2Vec implementation in Python. GRAPE's fast implementation of Node2Vec was particularly useful in the context of this project as the graph of collaborations created in the previous step contained 90,307,136 edges for first names and 253,679,354 edges for last names.

We computed the name embeddings on NYUAD's High-Performance Computing Jubail server, allocating 450 GB of memory and 20 cores. The default embedding dimension of 100 was used and the total number of epochs was set to 30. Due to time limits, the *max_neighbors* variable, describing the maximum number of randomly sampled neighbors to consider for single walks, was set to 50. Finally, the embedding of first names was calculated in 21h57 and the embedding of last names took 1 day and 15 hours.

4.3 Race Taxonomy

We used the US Census Bureau categorization which classifies individuals into five mutually exclusive race groups: White, Black, Hispanic, Asian, and Other. We removed the category "Other" from our taxonomy as it is an amalgam of Native American and other races, hence a heterogeneous

group. The choice of race instead of ethnicity was motivated by the similar race classification reported in the voters' dataset [14], which was used for the training, validation, and testing phases of our classifier. Moreover, ethnicity categories vary among regions and there is no general consensus as studies usually define their own ethnicity classification [4, 6, 7].

4.4 Generation of Synthetic Full Names and Preprocessing

The voters' dataset provides ground truth race data for first and last names separately. However, name-based race classifiers use full names as inputs to achieve good accuracy. Therefore, using two different methods, we gathered first and last names from the voters' dataset and generated two lists of synthetic full names.

In the first place, we selected all first and last names which were assigned a race with a probability greater than 90% and which could be given an embedding representation. Based on this selection, we created all possible full name combinations for White, Black, Hispanic, and Asian races. We randomly selected 1 million names for each race category leading to a dataset of 4 million full names. Then, we split it into a training set (80% or 3,200,000 rows), test and validation sets (each 10% or 400,000 rows). Finally, we selected 1000 names from the test set to create a benchmark set, which was later used to compare our results with Namsor.

The second method consisted of splitting the training, test, and validation sets before generating combinations. After having gathered all first and last names assigned to a race with a probability greater than 90%, we split the list by race and subsequently divided it into 3 sub-lists for training (80%), testing, and validation (10%). Then, we separately computed combinations for the three sub-lists and all four races. Finally, for each race, we selected 800,000 combinations from the training set and 100,000 combinations from the test and validation sets, leading to a training set of 3,200,000 full names and 2 smaller sets of 400,000 full names for test and validation phases. This method was performed to force the training, test, and validation sets to have no overlap. In other words, any first or last name in the test set could not be found in the training or validation sets.

4.5 The Classifier

We used the scikit-learn implementation of KNN (K-Nearest Neighbors), a supervised learning method that classifies a data point based on the K closest neighbors in the training data. After training the classifier, we used the validation set to find the number of neighbors giving the highest accuracy. The results were as follows:

- Training set with synthetic full names from method 1: 40 neighbors
- Training set with synthetic full names from method 2: 150 neighbors

5 EVALUATION OF NAME EMBEDDING

We evaluated the accuracy of the name embedding with the Olympics dataset. Although race is not a self-reported attribute, athletes usually represent their country of origin in sports. Hence, we gathered race information from the country athletes represented. In addition, we removed countries with a high immigration rate (United States, Canada, Australia, and countries in Western Europe) to increase confidence in the data.

In order to map country to race, we used the nationality taxonomy created by the authors of *Nameprism* [4]. This mapping was designed based on Cultural, Ethnic, and Linguist (CEL) similarities, represents 118 countries, and covers 90% of the world's population. Finally, we manually mapped ethnicity to race.

The choice of the Olympics dataset over the voters' dataset was motivated by the high diversity of athletes' names, the Olympics being an international event. In addition, the Olympics dataset is smaller (134,732 unique full names), making it easier to produce clear visualizations.

5.1 Race Representation

Name embedding was created using names in the MAG dataset, thus could not possibly cover all existing first and last names worldwide. In other words, a small proportion of first and last names could not be represented by a point in the embedding space, hence were left without any race prediction. We evaluated the representativeness of the embedding, making sure that it assigns an embedding representation to the same proportion of White, Black, Asian, and Hispanic names in the Olympics dataset.

Race	% of First Name	% of Last Name
White	90.7%	72.7%
Black	86.7%	76.6%
Asian	87.2%	63.9%
Hispanic	91.0%	89.2%

Table 1: Proportion of First and Last Names With Embedding

Table 1 gathers, for each race, percentages of unique first and last names from the Olympics dataset for which an embedding representation was available. The results show that the proportion of Hispanic first and last names with assigned

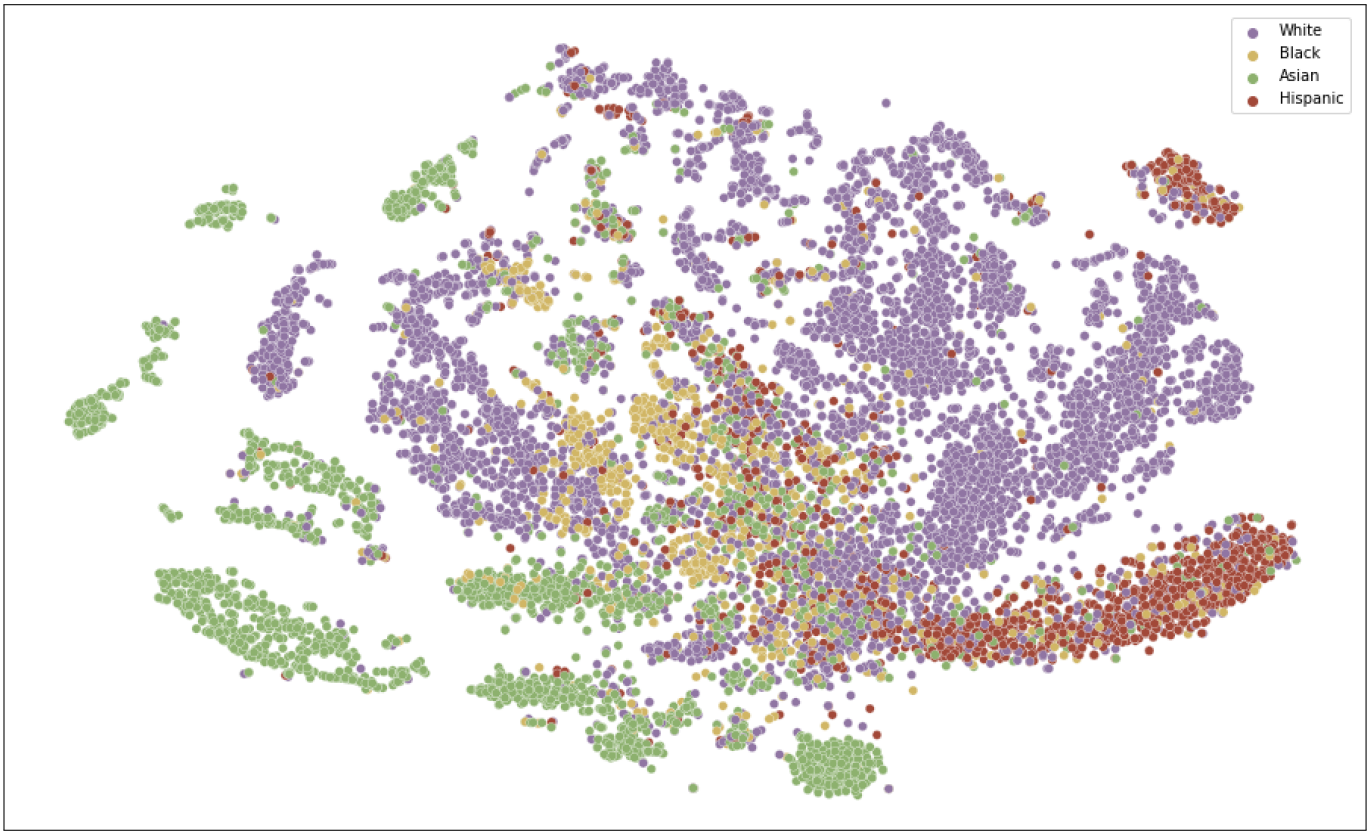

Figure 2: First Name Embedding Visualization

embedding is the highest among all four races. Overall, the embedding is fair and does not significantly give an advantage to a specific race category.

5.2 Name Embedding Visualizations

We used Principal Component Analysis (PCA) [19] and Distributed Stochastic Neighbor Embedding (t-SNE) [20] methods to project the 100D first and last name embeddings into 2D.

PCA is a method to reduce the dimensions of a dataset while retaining features that keep the maximum amount of variation and information about the data's original structure. We used this technique to reduce the initial embedding dimension from 100 to 50 and calculated that the retained features accounted for 98.7% of the variations in the first names dataset, while it reached 93.0% in the last name dataset.

We subsequently used t-SNE, a probabilistic technique for dimensionality reduction particularly well suited for the visualization of high-dimensional datasets. According to the authors of the original paper [20], t-SNE is computationally expensive, hence the need to use PCA first to "speed [..] up the computation of pairwise distances between the data points and suppress [..] some noise without severely distorting the interpoint distances" [20]. This method "minimizes the divergence between two distributions: a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding" [20].

The visualizations were produced for all first and last names associated with a unique race category in the Olympics dataset. Figure 2 shows the map of first names. We observe distinct clusters for each race category. However, while Asian, Hispanic, and White groups stand in distinct regions, the cluster for Black individuals is more dispersed. Figure 3 illustrates the landscape of last names: clusters are still visible, but less clear than on the map of first names.

Overall, the name embedding seems to be the most accurate for the Asian category (green). In particular, the very distinct green clusters (on the bottom left of the first name map and on the top left and bottom right of the last name map) mostly gather Japanese names, indicating that Japanese authors collaborate more with each other but with fewer international researchers. This "lack of international engagement with international researchers" [21] has indeed

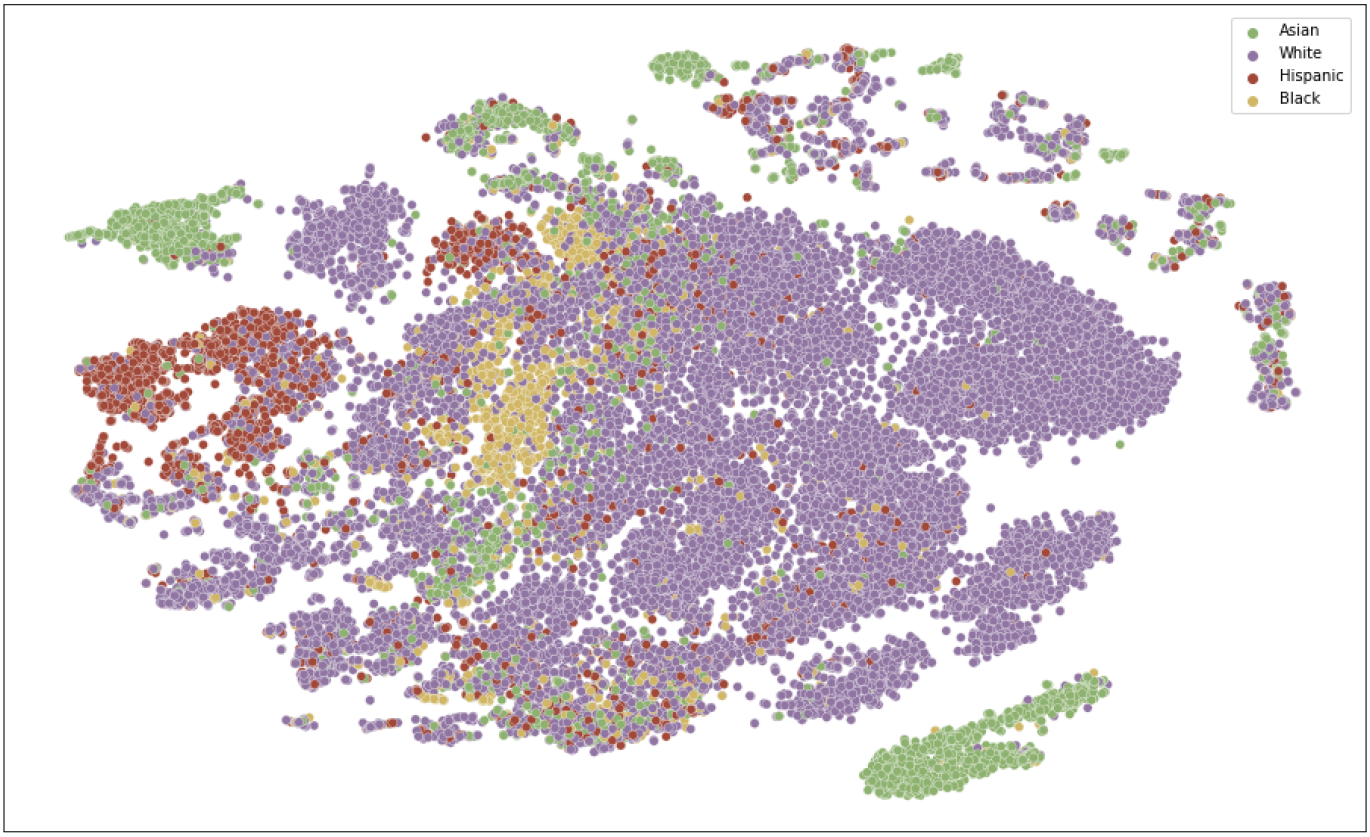

Figure 3: Last Name Embedding Visualization

been studied [21, 22]. Between 2003 and 2014, only 25% of all papers written in Japanese institutions were internationally co-authored, compared to 52% in the UK [22]. These visualizations highlight this homogeneity and strongly indicate that the name embedding captures race and nationality information well.

5.3 Quantitative Study: 10-Nearest Neighbors

Finally, we ran experiments to validate our observation quantitatively, evaluating the performance of the name embeddings with a nearest neighbor algorithm. We assigned a predicted race to each name based on the most frequent race among its 10 nearest neighbors in the embedding space. A score between 0 and 10 was also associated with the prediction, depending on the number of neighbors associated with the predicted race. Finally, we compared the prediction with the real race label.

The results for first names show an accuracy of 70% and a mean score of 8. In other words, 70% of first names were assigned to a race that was the most frequent race among their 10 nearest neighbors, and the average number of neighbors associated with the predicted race was 8 out of 10. Last names have an accuracy of 77% and a mean score of 8.5.

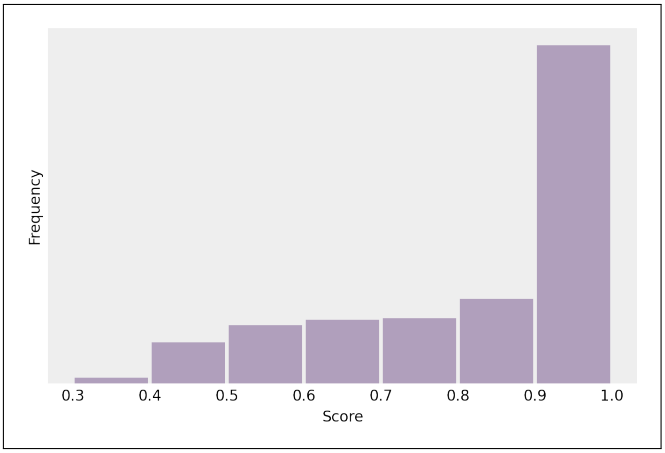

Figure 4: Distribution of Prediction Scores for First Names

Figures 4 and 5 display the distribution of scores among correct predictions. We observe that the majority of scores are close to 1, indicating a high level of confidence in the predictions. Overall, these results show a strong name embedding and highlight that using collaborations in academia is effective to capture the underlying racial and ethnic dimension of names.

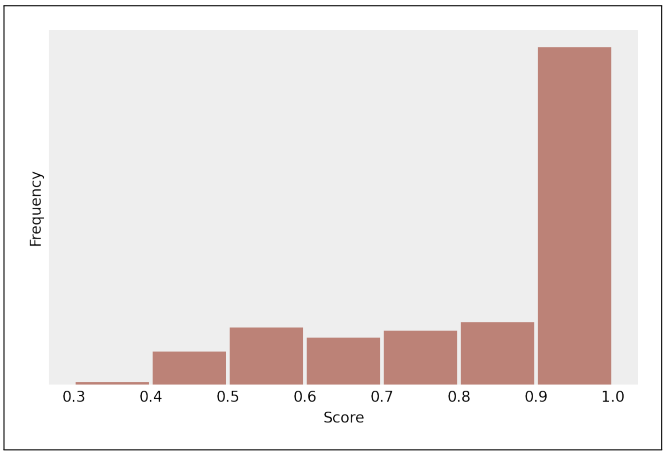

Figure 5: Distribution of Prediction Scores for Last Names

In addition, table 2 displays the 10 nearest neighbors of 5 popular first names from the UAE, China, Japan, Brazil and France and serves a qualitative analysis.

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

Table 2: Nearest Neighbors of Five Popular First Names

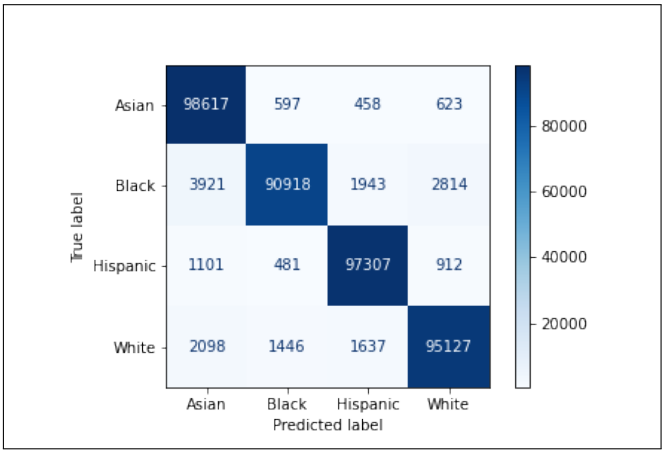

Figure 6: Confusion Matrix for the Test Set (Method 1)

6 RESULTS

As mentioned in section 4.4, the training, test, validation, and benchmark sets were created using two different methods. Method 1 consisted of generating full name combinations before splitting the dataset for the training, testing, and validation phases. Method 2 consisted of splitting the data before generating the combinations to ensure no overlap between training and test sets. We successively trained and tested the classifier with data from method 1 and method 2. In order to compare the results with *Namsor*, we had two different benchmark sets, extracted from the test sets produced by methods 1 and 2.

6.1 Method 1

The classifier is **95.49%** accurate on the test set with an F1 score of **0.955**. Regarding the benchmark set, the accuracy is **95.70%** with an F1-score of **0.957**. In addition, figures 6 and 7 display the confusion matrices for both sets.

The confusion matrices show that the results are fair across races, with the number of good predictions being approximately the same for each race. This observation is confirmed by analysis of the individual F1 scores for each race group on the test set:

- White: 0.952
- Black: 0.942
- Hispanic: 0.986
- Asian: 0.957

We subsequently passed the benchmark set to *Namsor* algorithm, which obtains an accuracy of **76.90%** and an F1 score of **0.766**. Therefore, this first experiment is successful as we obtain a remarkable accuracy of 95.49% on the test set and perform better than *Namsor* on the Benchmark set.

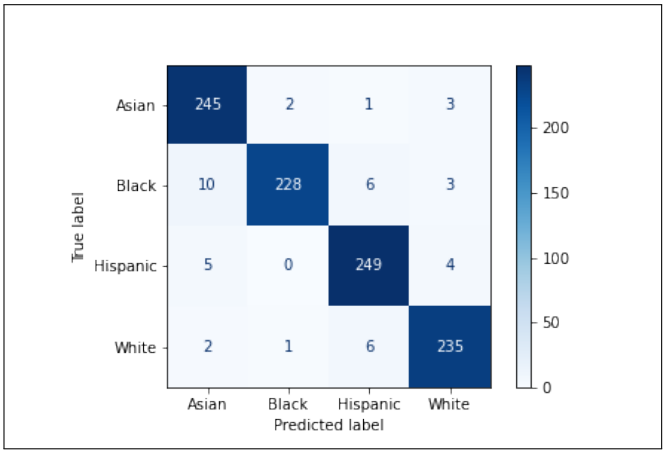

Figure 7: Confusion Matrix for the Benchmark Set (Method 1)

6.2 Method 2

The classifier obtained an accuracy of **70.28%** and an F1 score of **0.721** on the test set. Regarding the benchmark set, it is **72.50%** accurate with an F1-score of **0.723**. The confusion matrices are displayed in figures 8 and 9.

Figures 8 and 9 show that despite a lower accuracy compared to the first experiment, the predictions are still balanced among race categories, with individual F1 scores of:

- White: 0.652
- Black: 0.659
- Hispanic: 0.744
- Asian: 0.754

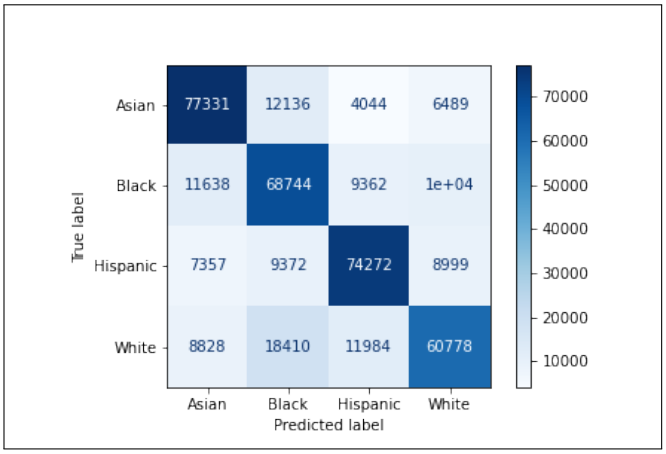

Figure 8: Confusion Matrix for the Test Set (Method 2)

We subsequently passed the corresponding benchmark set to *Namsor* algorithm, which obtains an accuracy of **77.80%** and an F1 score of **0.770**. For this experiment, *Namsor* performs better than our classifier.

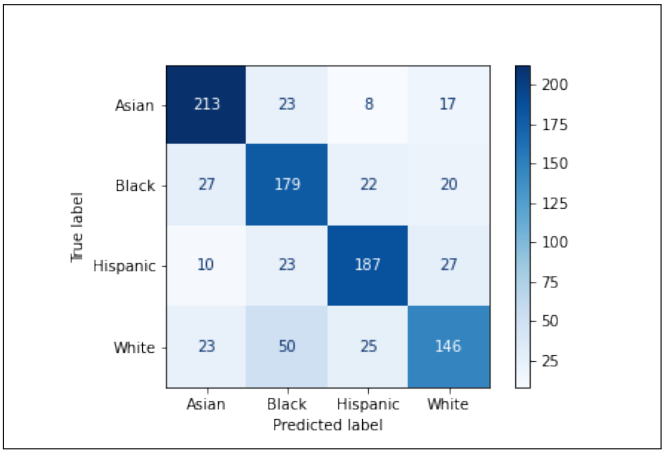

Figure 9: Confusion Matrix for the Benchmark Set (Method 2)

7 DISCUSSION

We trained and tested the classifier using training, validation, and test sets generated with two different methods. In general, the distribution of values in a test set, here names, is supposed to be as close to the real-world distribution as possible, to avoid over-fitting the training set. However, by doing the train-test split with method 1 (generating synthetic full names before splitting the data), the test set became very similar to the training set and not necessarily representative of the real world. In other words, the probabilities of a name appearing in the training and test sets were similar, but both higher than the probability of it appearing in the real world. This is due to the fact that we only selected names with strong race signals (with 90% confidence) from the voters' dataset. One attempt to solve this issue was to split the lists of first and last names before generating the combinations of synthetic full names (method 2). This forced the training, test, and validation sets to have no overlap and ensure there was no leakage problem.

While method 2 did not guarantee that the test set was closer to the real-world distribution of names, its lower accuracy and F1-score (70.28% and 0.721 compared to 95.49% and 0.955 for method 1) helped highlight bias in the first method that needs to be addressed with further research. Nonetheless, the results are not too far from those obtained with Namsor (77,80% accuracy and F1-score of 0.770), considering that KNN, a very basic method, was used. This highlights that the name embedding is overall very good, and that collaboration is a good factor to predict race.

Another objection to the methodology looks at the variety of full names for each race. To create the datasets, we randomly sampled the same number of names for White, Black, Asian, and Hispanic races. However, the dataset from which we sampled these names was larger for White names. This is due to the fact that the initial lists of first and last names were extensively bigger for the White category. Therefore, despite balanced predictions and an equal number of names for each race in the training data, the list of full names for Hispanic, Asian, and Black categories might have been less varied than for White names. This might have led to a classifier that correctly predicts a larger variety of White names than for Asian, Black, and Hispanic names.

Finally, one might say that Namsor results cannot be compared with this classifier's results as Namsor was created to predict race with first names, last names, and country of residence. However, we believe that the comparison is still necessary as most datasets of names do not contain information about the country of residence.

8 CONCLUSION

We demonstrated that using collaborations in academia 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. We further used the name embedding to build a new name-based race classifier, performing relatively well compared to the state-of-the-art race classifiers today, such as Namsor.

In addition, we addressed the problem of representation by producing an embedding that roughly covers the same percentage of White, Black, Hispanic, and Asian names. Moreover, we used training, validation, and test sets with the same number of synthetic full names from each race category. This balanced dataset was used for the classifier which achieved predictions with approximately the same accuracy and F1 score for each race group.

Despite the fact that this classifier had impressive results, there is still a need for more research and a few improvements can be suggested. First, the current method can only be used for names that are given a representation by the name embedding algorithm. Further research should be focused on including more names in the embedding database or using closest distance algorithms to associate similar names to the same embedding. Moreover, bias in the creation of synthetic full names should be studied. Furthermore, some language models will be valuable in the process of learning name embedding as they can extract substrings or features related to the structure of names that depend on race. Finally, other classifiers, such as random forests, support vector machines, or more complex deep learning models, should be investigated.

9 ACKNOWLEDGEMENTS

I express my deepest gratitude to Fengyuan "Michael" Liu for invaluable and constant support throughout the three 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] Donald A Barr. *Health disparities in the United States: Social class, race, ethnicity, and health.* JHU Press, 2014.

[2] Esteban González Burchard, Elad Ziv, Natasha Coyle, Scarlett Lin Gomez, Hua Tang, Andrew J Karter, Joanna L Mountain, Eliseo J Pérez-Stable, Dean Sheppard, and Neil Risch. The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine*, 348(12):1170–1175, 2003.

[3] Kevin Lang and Ariella Kahn-Lang Spitzer. Race discrimination: An economic perspective. *Journal of Economic Perspectives*, 34(2):68–89, 2020.

[4] Junting Ye, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, and Steven Skiena. Nationality classification using name embeddings. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1897–1906, 2017.

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

[6] Junting Ye and Steven Skiena. The secret lives of names? name embeddings from social media. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3000–3008, 2019.

[7] Vetle I Torvik and Sneha Agarwal. Ethnea–an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database. 2016.

[8] Fangzhou Xie. rethnicity: An r package for predicting ethnicity from names. *SoftwareX*, 17:100965, 2022.

[9] Lokman I Meho. Gender gap among highly cited researchers, 2014–2021. *Quantitative Science Studies*, pages 1–28, 2022.

[10] Anthony J Videckis, Alisa Malyavko, Denver B Kraft, and Sean A Tabaie. Male versus female authorship in flagship pediatric orthopaedic journals from 2002 to 2021. *Journal of Pediatric Orthopaedics*, pages 10–1097, 2023.

[11] Sherri Lynn Conklin, Michael Nekrasov, and Jevin West. Where are the women: The ethnic representation of women authors in philosophy journals by regional affiliation and specialization. *European Journal of Analytic Philosophy*, 19(1):S4–48, 2023.

[12] Paul Sebo. Publication and citation inequalities faced by african researchers. *European journal of internal medicine*, 106:135–137, 2022.

[13] Mengchen Sam Yong, Lavínia Paganini, Huilian Sophie Qiu, and José Bayoán Santiago Calderón. The diversity-innovation paradox in open-source software. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*, pages 627–629. IEEE, 2021.

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

[15] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*,

27(1):415–444, 2001.

[16] Martin Atzmueller and Florian Lemmerich. Homophily at academic conferences. In *Companion Proceedings of the The Web Conference 2018*, pages 109–110, 2018.

[17] Luca Cappelletti, Tommaso Fontana, Elena Casiraghi, Vida Ravanmehr, Tiffany J Callahan, Marcin P Joachimiak, Christopher J Mungall, Peter N Robinson, Justin Reese, and Giorgio Valentini. Grape: fast and scalable graph processing and embedding. *arXiv preprint arXiv:2110.06196*, 2021.

[18] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

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

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

[21] Shannon Mason. Adoption and usage of academic social networks: a japan case study. *Scientometrics*, 122(3):1751–1767, 2020.

[22] Suvendrini Kakuchi. Restrictions on collaboration are 'hindering' international research. *University World News*.