Assigning Nationality to Names Using Machine Learning to Differentiate Emigration from Return Migration of Scholars

Faeze Ghorbanpour[1,2*] and Aliakbar Akbaritabar[2*]

[1]Ludwig Maximilian University of Munich, Munich, Germany
[2]Department of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany.
(*Corresponding authors: ghorbanpour@demogr.mpg.de, akbaritabar@demogr.mpg.de)

Current version: November 1, 2023

Abstract

Most digital trace data does not include the nationality of individuals for privacy reasons. Once this data is used for migration research, it can have a left truncation issue since we are uncertain about the migrant's country of origin. Identifying nationality enables a better differentiation between emigration and return migration. We detect the nationality with the least available data, full names, and use it instead of the country of academic origin in studying the migration of scholars. We gathered 2.6 million unique name-nationality pairs from Wikipedia and categorized them into families of nationalities with three granularity levels. We used a character-based machine learning model that reached a weighted F1-score of 80% for highest- and 64% for country-level categorization. We discuss the shifts in migration rates when considering the assigned country of origin based on authors' names rather than the previously used country of first academic affiliation. Our results show that this impact is exacerbated in the case of countries of immigration that have a more diverse academic workforce such as the USA, Australia, and Canada.

Keywords: Machine Learning; Bibliometric Data; Migration of Scholars; Nationality

Introduction

Sociological and demographic research increasingly uses digital trace data (*1*, *2*). The nationality of individuals is not included in most of these digital traces to preserve privacy. While using only observational data in migration research, we cannot be certain about the country of origin of the migrants. This causes the migration trajectories built using the digital traces to be prone to a left truncation issue. This is because we do not know if the first emigration event observed in digital traces is from the actual country of origin or if it is recorded in a next location along the migration trajectory. Hence, there is a need to identify the nationality of people with the least available information, their names. An accurate name-based nationality identification complements observational data by adding a baseline for comparison to resolve such uncertainties in migration trajectories.

Bibliometric data, which is information extracted from scientific publications (*1*), is a valuable resource for analyzing the migration patterns of scholars (*3*–*5*). Bibliometric databases do not provide information about a scholar's country of origin or nationality. As a result, researchers utilize various methods to determine the country of origin for these individuals, such as using census data (*6*, *7*) or using the country where a scientist first published their initial paper as their country of academic birth (*1*, *3*, *8*). We aimed to provide an alternative method for identifying scholars' nationalities through bibliometric data. By recognizing the nationalities of scholars, the composition of the population of scholars per country and rates of migration could be updated with more nuanced measurements. Furthermore, it could help explain some of the observed trends in the migration of scholars as *return migration* of the graduate students to their home countries instead of emigration (*5*).

Machine learning (ML) models excel in text-based tasks, recognizing patterns in both common and uncommon names. A nationality detection system that uses ML is highly scalable, enabling quick analysis of large datasets. The names of people have been used in many studies to detect nationality, ethnicity, and race using machine learning methods. Lee *et al.* (*9*) and Le *et al.* (*10*) use recurrent neural networks to identify the nationality of individuals based on their names, while Kang (*11*) proposes a convolutional neural network classification method for name-nationality classification. However, these studies only use a limited amount of data, and their analysis is solely based on the training data, which can introduce bias in the results. To minimize bias towards the dataset, it is important to train the machine learning model using as large as possible amount of data and to evaluate the model on a dataset that is different from the training dataset.

Nameprism (*12*, *13*) presents a name embedding that utilizes Naive Bayes to classify nationalities. Their model is able to show the similarities and dissimilarities of names between nationalities with the distance between the embeddings of people's names. However, the inaccessibility of the data and code restricts other researchers from modifying nationality categories or conducting additional research.

Our paper seeks to analyze individuals' names to determine their nationality and ethnicity. Our study compiled a dataset of 2.6 million unique name-nationality pairs of individuals who have a page on Wikipedia. Our training data is more extensive than previous research, which relied on public data. We will publicly share the developed models, countries' categorization, and gathered data to facilitate replications and extensions of our study.

Data and Methods

A dataset of pairs of names and nationalities was required to train our classification models. Full names serve as our textual data, and nationalities serve as their classes. Once our model is trained well, and for the

analysis step, only the names of scholars from Scopus are sufficient to identify the origin of the name and conduct an analysis of their country of origin.

Individuals who possess a Wikipedia page can serve as a valuable resource for acquiring our dataset. These pages typically provide information regarding a person's nationality, city of birth, or citizenship. We conducted an extensive search for publicly available datasets sourced from Wikipedia. We found and combined seven distinct datasets such as Wikipedia notable people's mobility (*14*) and WikiBio (*15*). We aimed to study the changes in scholarly migration patterns and rates. Therefore, we have utilized a dataset acquired from Scopus. This dataset comprises a vast collection of 17 million full names of scientific article authors, along with the country in which they are conducting their research.

Following Le *et al.* (*10*) and Ye *et al.* (*12*), we have developed a hierarchical three-level classification model of nationalities. However, we modified these categories to fit our data better. These categories range from general groups of families of nationality or ethnicity with ten classes to a more specific country level with 180 classes. Each level corresponds to the categorization of the lower level. This approach allows for a clearer understanding of the similarities and closeness of names among individuals from different countries.

Fig. 1: Treemap of countries categorization in three levels from inner (lower granularity) to outer (higher granularity) circles.

Figure 1 presents a treemap of our countries' categorization. It is important to note that certain countries, including the United States, Canada, Australia, New Zealand, and South Africa, are characterized by a more diverse and multicultural population. Consequently, similar to previous frameworks (*10*, *12*), we excluded these countries from our training and testing steps and used them in the evaluation step. Additionally, this figure shows the distribution of classes at each level; level 1's largest class is German, and level 3 is British.

The letters in the names determine the patterns in the names, so we used character-based machine-learning methods to detect these patterns. Currently, we have developed and tested two machine-learning

models: FastText (*16*) and a character-based Convolutional Neural Network (CNN) (*17*). FastText employs a skip-gram model approach, representing each word as a collection of character n-grams. This approach considers subword information, contributing to a more comprehensive representation. The character-based CNN comprises multiple layers of Convolutional Neural Networks specifically designed for text classification purposes. Given the imbalanced nature of our data, we selected two evaluation metrics: weighted F1-score and balanced accuracy. By utilizing these evaluation metrics, we aimed to obtain a comprehensive understanding of the performance and effectiveness of our models in handling imbalanced data.

Preliminary Results

Table 1 provides a comparison of the performance of our models on the test set derived from the Wikipedia dataset. The results demonstrate that FastText outperforms the character-based CNN across all three levels of classification in terms of accuracy and F1 score. At the first level of classification, the model achieves an accuracy of approximately 81%. At the second level, the accuracy remains high at around 78%. However, as the number of classes increases at the third level of classification, the model's performance decreases to 64%. Nevertheless, considering the complexity arising from the 180 classes at this level, this performance can still be considered acceptable. Nonetheless, there remains an opportunity to enhance the performance of the model, particularly when dealing with a larger number of classes at the third level.

Table 1: Comparison table of the utilized machine learning models.

	Level 1		**Level 2**		**Level 3**	
	Balanced Accuracy	**Weighted F1-score**	**Balanced Accuracy**	**Weigted F1-score**	**Balanced Accuracy**	**Weigted F1-score**
Char CNN	72.1	74.5	60.9	65.1	40.0	43.5
Fasttext	81.3	81.4	78.0	77.6	64.7	62.5

We have examined the model's performance for each class at level 1 classification as shown in Figure 2. The assessment criteria used are precision, recall, and f1-score. It is evident that the model demonstrates higher precision compared to recall, indicating that the model accurately predicts the cases for each class. Additionally, for classes with more data, such as European regions, the average f1-score exceeded 80%, indicating the model is more effective at detecting European names and adding more training data could in principle increase the performance for other classes.

Figure 3 compares the country of academic birth (i.e., the country of affiliation in the first publication) using Scopus data with our assigned nationality using the author's full name for eight level 2 regions (top panel). It is evident that authors from non-European regions such as those with Chinese, Japanese, Persian, and Indian names are more likely to publish their first paper in an institution located in these countries. In other words, their country of academic birth and ML-assigned nationality are similar in most cases (above 70%). In contrast, authors with first publications in institutions located in European regions such as German, French, English and Russian are more diverse in terms of the composition of authors based on their ML-assigned nationality using names. In most of these cases close to or above 50% of authors have origins that are different from their academic origin. These could be higher education graduates who are originally from elsewhere studying in migrant-receiving countries with more advanced higher education systems.

The bottom panel of Figure 3 compares the country of academic origin and ML-assigned origin of authors for countries that were excluded during the training process due to their more diverse scientific workforce such as the USA, Australia, New Zealand, and Canada. It is clear that authors who published their first publication in these countries have a wide variety of ML-assigned countries of origin. As an example, only

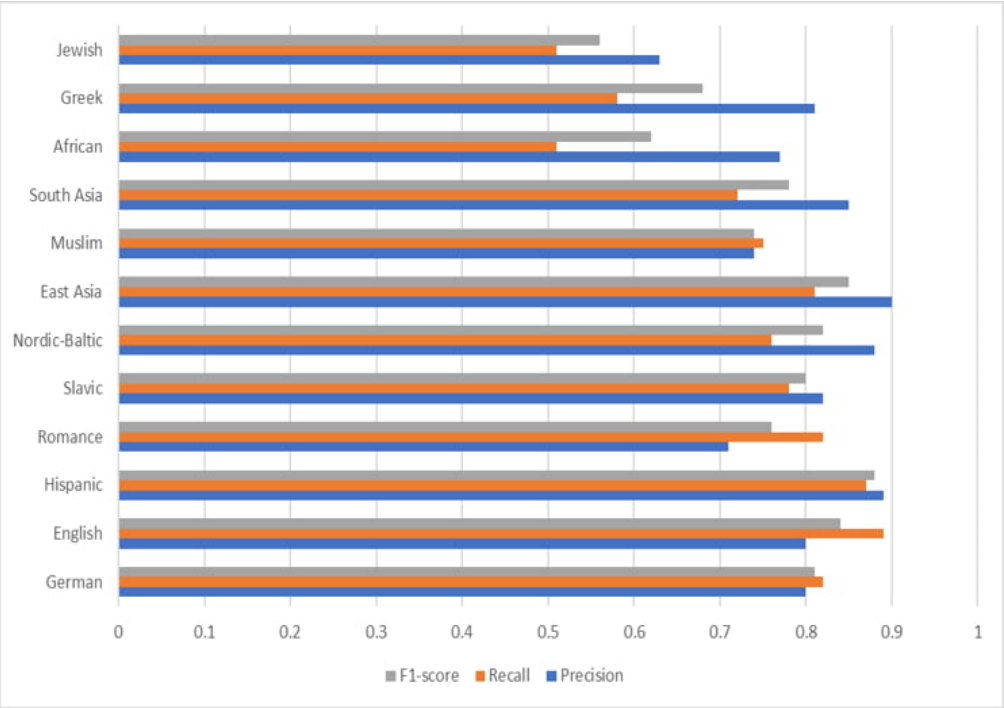

Fig. 2: Evaluation metrics for the first level classification.

24% of authors affiliated with the USA institutions have names with English origins in contrast to 7% with Korean, 16% with Chinese, 12% with German, and 6% with Indian names. Of course, these could be second-generation migrants in the USA who have names attributed to their country of origin, still, using only the country of academic origin, we would consider them as Americans while these results highlight the diversity of the scientific workforce in the USA.

Conclusion

This study presents a supervised Machine Learning system that focuses on identifying nationalities based on people's names. The study demonstrates a substantial improvement by providing a wider coverage of countries. Furthermore, the data collection process was enhanced, with a substantial increase in the number of names and nationalities gathered from Wikipedia (2.6 million pairs of names and nationalities). The model achieves almost 80% f1-score for both level 1 and 2 classifications. Additionally, the model exhibits an 80% precision rate in returning data despite lower performance in recalling data for small classes.

Our results have important implications for migration research using digital trace data (*1*) and specifically studies on scholarly migration. Using our ML-assigned country of origin could add more nuance to the previously used country of academic origin (*3–5*). In the context of the migration of scholars using bibliometric data, a left-truncation issue is present. This issue relates to the fact that we are not sure if the publication trajectory has started before the starting date covered in the bibliometric data. Using the country of affiliation in the first publication has been the sole source of metadata to define from where the migration trajectory starts (*1*). Our results help complement and add more nuance to this picture by using the ML-assigned country of origin in comparison to the country of academic origin. Of course, part of the observed diversity in the composition of the scientific workforce relates to the second-generation migrants

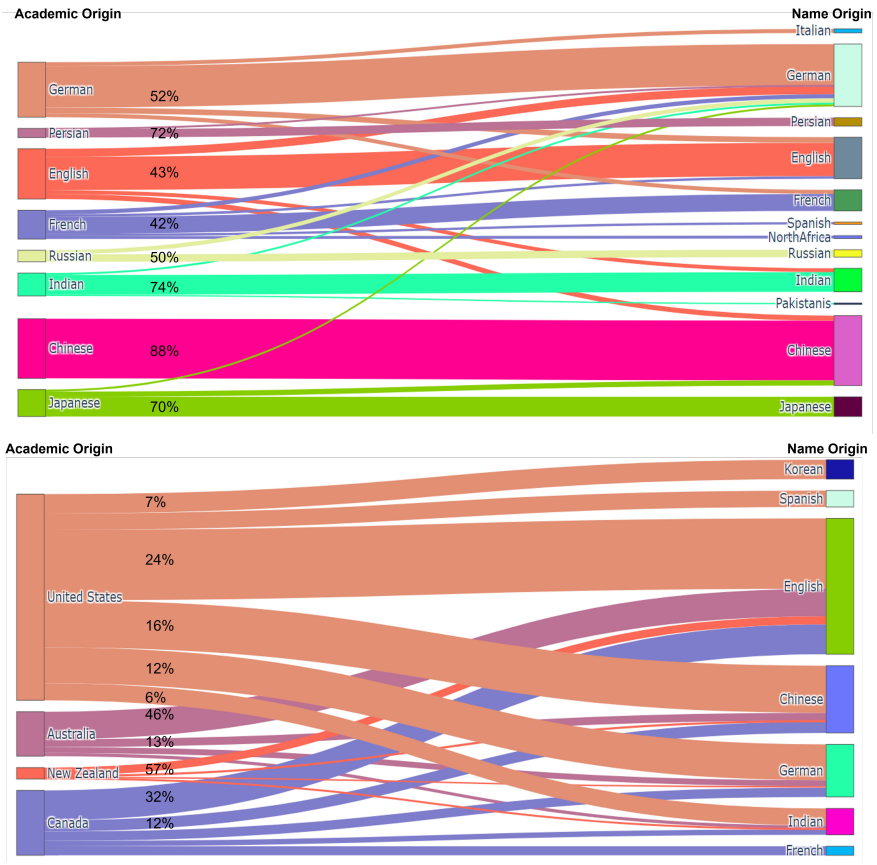

Fig. 3: Comparing the country of academic origin (left column) and assigned nationality using ML model (right column) at level 2 regions (top panel) and for the case of so-called immigration countries which have a diverse scientific workforce (bottom panel) that were excluded from training data.

who have names attributable to their country of origin. Once we observe that some of these scholars emigrate from certain countries, using the ML-assigned countries, we can differentiate between "emigration" and "return migration" which was previously not possible (*5*) by using only academic origin.

In addition to the presented improvements, we acknowledge certain limitations, including the challenge of dealing with a high number of classes and an imbalanced dataset in country-level classification. The significant improvements in coverage, data collection, and model performance have established a solid foundation for future research in this domain. By obtaining data from other platforms and utilizing advanced algorithms, further enhancements will be achieved to ensure more accurate and comprehensive nationality detection. Furthermore, the rates of migration of scholars between bilateral pairs of countries need to be calculated to show the impact of our proposed methodology and ML-assigned country of origin on aggregate rates of scholarly migration and academic brain circulation worldwide.

Acknowledgements

We thank our colleague Tom Theile for lending us one of the seven training datasets.

References

1. R. Kashyap *et al.*, in *Research Handbook on Digital Sociology* (Edward Elgar Publishing, Mar. 2023), chap. Research Handbook on Digital Sociology, pp. 48–86, ISBN: 978-1-78990-676-9.
2. D. Alburez-Gutierrez *et al.*, Publisher: SocArXiv (2019).
3. X. Zhao, A. Akbaritabar, R. Kashyap, E. Zagheni, *Proceedings of the National Academy of Sciences* **120**, e2214664120 (Mar. 2023).
4. A. Akbaritabar, T. Theile, E. Zagheni, "Global flows and rates of international migration of scholars", tech. rep. (Max Planck Institute for Demographic Research, Rostock, Germany, 2023).
5. E. Sanliturk, E. Zagheni, M. J. Dańko, T. Theile, A. Akbaritabar, *Proceedings of the National Academy of Sciences* **120**, e2217937120 (2023).
6. G. Lewison, P. Roe, R. Webber, R. Sullivan, *Scientometrics* **106**, 105–117 (2016).
7. J. Grilli, S. Allesina, *Proceedings of the National Academy of Sciences* **114**, 7600–7605 (2017).
8. M. Thelwall, *Journal of Data and Information Science* **8**, 1–25 (2023).
9. J. Lee *et al.*, presented at the IJCAI, vol. 17, pp. 2081–2087.
10. T. T. Le, D. S. Himmelstein, A. A. Hippen, M. R. Gazzara, C. S. Greene, *Cell Systems* **12**, 900–906 (2021).
11. Y. Kang, presented at the Proceedings of the 2020 2nd international conference on big data engineering, pp. 70–74.
12. J. Ye *et al.*, presented at the Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1897–1906.
13. J. Ye, S. Skiena, presented at the Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3000–3008.
14. L. Lucchini, *Wikipedia notable people's mobility*, version V1, 2019, (https://doi.org/10.7910/DVN/PJS21L).
15. R. Lebret, D. Grangier, M. Auli, presented at the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP).
16. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, *Transactions of the association for computational linguistics* **5**, 135–146 (2017).
17. X. Zhang, J. Zhao, Y. LeCun, *Advances in neural information processing systems* **28** (2015).