Wikidata as a Source of Demographic Information

Samir Abdaljalil*
Texas A&M University
College Station, TX, USA
sabdaljalil@tamu.edu

Hamdy Mubarak
Qatar Computing Research Institute
Doha, Qatar
hmubarak@hbku.edu.qa

Abstract

Names carry important information about our identities and demographics such as gender, nationality, ethnicity, etc. We investigate the use of individual's name, in both Arabic and English, to predict important attributes, namely country, region, gender, and language. We extract data from Wikidata, and normalize it, to build a comprehensive dataset consisting of more than 1 million entities and their normalized attributes. We experiment with a Linear SVM approach, as well as two Transformers approaches consisting of BERT model fine-tuning and Transformers pipeline. Our results indicate that we can predict the gender, language and region using the name only with an accuracy of over 0.65. The country attribute can be predicted with less accuracy. The Linear SVM approach outperforms the other approaches for all the attributes. The best performing approach was also evaluated on another dataset that consists of 1,500 names from 15 countries (covering different regions) extracted from Twitter, and yields similar results. We share the datasets used in our experiments in addition to an online interface for testing and API calling.

1 Introduction

Names are an important part of humans' identities. They can act as a representation of a lot of characteristics that an individual can possess (Wilson, 2003). According to Mbarachi and Igwenyi (2018): "Naming is one of the practices that emphasize the mutual relationship that exists between language and culture, wherein culture embodies language, while language expresses culture." This makes human names highly interesting for the potential information they carry. Many factors, including the country, language, and dialect of a person, could influence the way a name is written, and the character combinations used within that name (Abdulganiy et al., 2015). Given the name "Jakob", which could be categorized as a Western European name, when surveying North American context, another version "Jacob" is more frequent and dominant. Despite being considered somewhat private and not made public, having access to such data can help in many contexts, including linguistic studies and user demographic research (Jung et al., 2021). In addition, predictions made based on names can relieve users from tedious form filling, and help to better tailor user preferences and provide more precise suggestions for attributes like gender, nationality, preferred language, etc.

In this paper, we investigate predicting the country (and region), gender, and language of an individual using one piece of information - their name. We are able to predict this information by training models using publicly available data from Wikidata (Refer to Appendix A for a demo). We also share a comprehensive dataset consisting of more than 1 million entities with their corresponding normalized country, gender, language, and region information[1]. This dataset consists of entities from a total of 100 countries covering 17 regions, 48 languages, and 2 genders in both Arabic and English.

The main contributions of this paper are: **(i)** We introduce a novel idea of using Wikidata to collect demographic information from names. We discuss challenges and details for data curation and normalization in building a comprehensive dataset for entities' names as well as their corresponding country, language, region and gender information; **(ii)** We perform experiments on both Arabic and English names, to demonstrate that our methodology is able to be generalized to more than one language; and lastly, **(iii)** We benchmark machine learning and BERT models on predicting attributes from person names using different evaluation sets.

*The contribution was made while the author was at Qatar Computing Research Institute.

[1] https://alt.qcri.org/resources/WikidataDemographics.zip

Proceedings of The Second Arabic Natural Language Processing Conference, pages 1–10

2 Related Work

Over the years, researchers have investigated the role names can play in deducing other information, such as gender, country and religion, about individuals. Such research requires a large amount of diverse names to be able to explore, which is why a large portion of the research used social media platforms as their primary source of data, namely Twitter (Ye et al., 2017; Huang et al., 2014; Goudarzi, 2020), as well as Facebook (Tang et al., 2011) and InterPals (Jung et al., 2021), while others used governmental and census data to gain access to individuals' information (Chaturvedi and Chaturvedi, 2022; Ye and Skiena, 2019; Kowsher et al., 2020; Hofstra and de Schipper, 2018). Both types of data come with their own sets of limitations, as social media can be noisy and unstructured, while governmental census data does not include a large amount of diversity in terms of nationality, as they tend to report on a national level.

Tzioumis (2018) focused on names from the United States, by creating a list of approximately 4000 first names and 11000 surnames and tagging them with their corresponding ethnicity through analyzing a combined dataset that consisted of mortgage application data. In addition, Huang et al. (2014) extracted Twitter user profiles based in Qatar to infer users' country based on features of their profile such as timezone, description, as well as their name. By using Gradient Boosted Tree, introduced by Friedman (2002), Huang et al. (2014) achieved an overall accuracy of 83.8%. Tang et al. (2011) used a list of 1.73 million Facebook profiles of individuals based in New York City, and applied a Multinomial Naive Bayes classifier, which led to an accuracy score of 95.2%. However, these investigations are somewhat limited in that they are focused on collecting names from specific countries. As a result, we address this issue by including large number of countries in our exploration, to ensure that there is both diversity and coverage in the results obtained.

Lee et al. (2017) built a recurrent neural network that predicts the country of an individual based on their name, by looking at character embeddings. They trained the model using 17,721 names and their corresponding country from the Olympics official website. In doing so, Lee et al. (2017) achieved an accuracy score of 51.9%. Malmasi (2014) developed their own dataset, The MQ Names Corpus, to investigate names and their gender and ethnicity classification. This dataset "contains over 13k names from 5 cultural groups...These include names of Arabic, German, Iranian and Japanese origin. Romanized versions of all names are used" (Malmasi, 2014). By employing a linear Support Vector Machine (SVM) classifier, character tri-grams achieved the best accuracy scores for both gender and ethnicity classification, which were 90% and 83%, respectively. Goudarzi (2020) investigated the relationship between the first name of an individual and their gender, by extracting twitter users' first names and assigning them a gender through using Amazon Mechanical Turk (AMT), which "Is a platform developed by Amazon for the distribution of short, simple tasks to legions of human workers around the world" (Goudarzi, 2020). This led to them creating a dataset of 12,681 name and gender pairs. They found that using an individual's name resulted in a 20% increase in the accuracy of the gender classifier when compared to a standard baseline classifier. Although these investigations yielded promising results, the main limitation is that they use relatively small datasets to train and evaluate their approaches.

Jung et al. (2021) collected approximately 1 million unique names from social media network InterPals, with their gender, country and age information. They then created an algorithm that takes the three aforementioned attributes as parameters, and assigns a demographically suitable name by using the data collected from InterPals. They opted for a manual evaluation of the names by ethnically-diverse evaluators, and achieved a score of 85.6%. Chaturvedi and Chaturvedi (2022) focus on South Asian names and their relationship with religion. Using data collected by the National Council of Applied Economic Research, which consisted of approximately 150,000 names, they found that an SVM taking into consideration both first and last names evaluated the highest, achieving an F1 score of 82.88%. Furthermore, using 89 million user profiles from Twitter, Ye and Skiena (2019) investigated what names infer about religion, gender and ethnicity, by using the profiles' information, paired with census data, to provide reference labels for each investigated attribute. Their analysis suggested that using embedding representations for the names yielded better results, when compared to substrings.

In addition to some of the limitations mentioned previously, we observed that, collectively, the re-

search in the field has been geared towards a subset of the traits explored in this paper: Country, Gender, Language and Region. To our knowledge, however, there is no investigation geared towards experimenting with what a name infers about all four attributes. Furthermore, although there was some exploration of this topic in diverse regions and/or countries, a large portion of the research used English versions of all names, regardless of origin. As a result, we are inspired by these investigations and take them one step further to investigate names, written in both English and Arabic, and what they infer about the Country, Gender, Language and Region.

3 Data Collection

3.1 Wikidata

Wikidata is a collaboratively edited multilingual knowledge graph hosted by the Wikimedia Foundation[2]. Wikidata is a common source of open data that Wikimedia projects can use such as Wikipedia (Vrandečić and Krötzsch, 2014).

Wikidata focuses on **items**, which represent any kind of topic, concept, or object.[3] Each item has the following attributes: label (ex: "Kofi Annan"), unique identifier (aka QID), description, and optionally aliases. Normally, these attributes are written in multiple languages. In addition, an item consists of "Statements" (in the form of "Property" and "Value") which show how any information known about the item is recorded in Wikidata. Figure 1 shows an example for the entity "Kofi Annan" as well as the various fields and information recorded in Wikidata.

We extracted person names from Wikidata dump on 2021-10-27[4] by filtering only entities having "**instance of**" (Property:P31) equals to "**human**" (Value:Q5). As we were interested in studying names written in languages from different families, we extracted names written in English and/or in Arabic for our study,[5] and we leave working on more languages for future work.

In this work, we focus on inferring the following demographic information from a person name: **gender** (property "sex or gender" - P21), **country** (property "country of citizenship" - P27), and **language** (property "language spoken" - P1412). This can be extended to other properties like **religion**[6] (property "religion or world view" - P140), **age** (property "date of birth" - P569), **ethnicity** (property "ethnic group" - P172), etc. which shows the usefulness of using Wikidata for other attributes.

4 Data Analysis

4.1 Original Data

The original data extracted from Wikidata consisted of 9,284,039 unique samples. Each entry consisted of the name in English, two different forms of Arabic (standard and Egyptian dialect), as well as the extracted Country, Gender, and Language information. Any missing information was replaced with *XX* for easy identification.

4.2 Data Cleaning & Normalization

We observed that many names of individuals were provided in various transcription format such as phonetic for *Muhammad ibn Jãbir al-Battãnĩ*. Furthermore, we found that categories, such as Country, had many classes due to historical or political changes. There were 2558 unique countries in English, and 1866 countries in Arabic. This is due to the fact that a lot of the data was labelled using different name variations for the same country or area. For instance, *Grazia Deledda* was classified under the country of *Kingdom of Italy*, while *Lorenzo di Credi* was classified under *Republic of Florence*. Although both individuals are from the modern country, Italy (IT)[7] but to reflect different time or era, some entries in Wikidata either state the specific city that an individual is from, or provide a different name or title for the country. The distribution of countries in the original data, excluding rows with *XX*, was highly imbalanced. For example, 55.4% of the data, belongs to only the top 15 countries, two of them being *Ming Dynasty* and *Qing Dynasty*, which are historical entities in what is currently as China (CN). This is illustrated in Appendix B. As a result, normalization was an essential part in ensuring that the data can be used efficiently.

We identified a list of normalized classes for each of the explored attributes to organize the data

[2] https://wikimediafoundation.org/

[3] https://en.wikipedia.org/wiki/Wikidata

[4] Wikidata dumps can be downloaded from: https://dumps.wikimedia.org/wikidatawiki/entities/

[5] In this version, names start with Q23:"George Washington" and end with Q108439676:"Ivan Gonzalvez Tamarit"

[6] Religion information was available for 173,379 entities.

[7] We use ISO 2-alpha for country codes: https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2

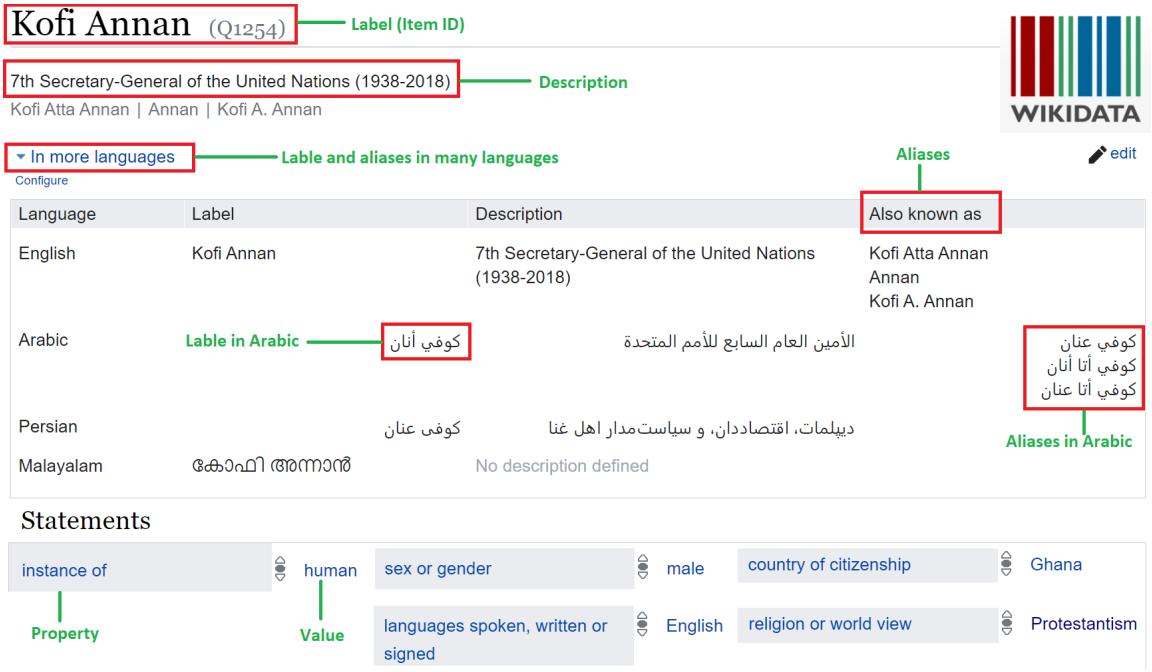


Figure 1: Wikidata sample page

appropriately and address any unnecessary variations within the data.

For **gender**, we assigned the following values: male and female. The full list of original values for gender and their mappings can be found in Appendix D.

For **country**, we chose 100 countries based on common factors such as their influence[8] and their GDP and population[9]. We plan to extend this work to cover all the World countries.

For **region**, we hypothesize that for some applications, fine-grained country classification might not be needed (ex: Ghana versus Ivory Coast), and coarse-grained geographical information (e.g. region) can be considered as satisfactory for many users (e.g. Western African region).

In addition, predicting exact country of origin of a person from their name can be less accurate as many names are common across neighboring countries. We added region information for the selected countries based on UN data[10] which summed to a total of 18 regions. Training a separate classifier to predict region of origin for a person name can be used to increase the confidence in country classification as we have many countries to predict. If country and region classifiers agree, this can further confirm that the accuracy in country prediction is high and vice versa.

For **language**, we chose the first official language of each of the selected counties,[11] and this summed to 48 languages. Please refer to Appendix E for more details about countries, regions, and languages.

Basic statistics about the selected attributes before and after class normalization are listed in Table 1.

Once the data was normalized, in order for us to take advantage of both the Arabic and English versions of each name, we considered each of them as a separate entry, alongside their corresponding normalized information.

In addition, we normalized the names by removing accents and titles. And for Arabic, we replaced some letters, namely أ، إ، آ، ة، ى with ا، ه، ي as they are used interchangeably by many users. We also replaced decorated letters (mainly imported from Farsi) with their plain equivalent letters (ex: letter ڤ to ف). An example is shown in Figure 2.

After performing data cleaning and normalization, and removing any duplicate names, we were

[8] https://www.usnews.com/news/best-countries/most-influential-countries

[9] https://worldpopulationreview.com/country-rankings/most-powerful-countries

[10] https://esa.un.org/MigFlows/CountryList.html

[11] https://en.wikipedia.org/wiki/List_of_official_languages_by_country_and_territory

left with a total of 8,339,570 entries. In table 2, we show the sizes of the data after normalization and cleaning. It is important to note that, since we consider different language variations of the same name as separate entries, for analysis in table 2, we consider each entity as one entry, regardless of whether they are included more than once in both Arabic and English. As a result, the total unique entities in the data amount to 7,462,850.

Attribute	Org. Classes	Norm. Classes
Country	2557	100
Gender	35	2
Language	974	48
Region	NA	18

Table 1: Unique classes in each attribute before (Org. Classes) and after normalization (Norm. Classes)

François-René de Chateaubriand —> francois rene de chateaubriand
جوستاف ليوبولد بليت <— جوستاف ليوپولد پليت

Figure 2: An example of two entities, in both Arabic and English, before and after cleaning.

4.3 Balanced Data

Since the normalization step was mainly for cleaning some of the unnecessary noise and variations that came from the original data. A major imbalance in some of the classes within the data, especially in the attribute with the most classes, country, was still apparent. To ensure that the data was appropriately balanced, we used the country attribute as a basis for our balancing. After experimenting with different ranges, we found that including a maximum of 20,000 entries per country resulted in a large enough dataset, while maintaining the balance between the classes as much as possible.

After balancing, the top 15 countries went from dominating the data at 54.6%, to 26.3%, which can be seen in Appendix C. In doing so, we ended up with a total of 1,096,910 entities, out of which 834,902 were unique entities. Total entities came from considering the different translations of the same name, and unique entities are based on their original QID from Wikidata.

Table 2 showcases the available entities corresponding to each attribute after normalization and balancing, in relation to the original data extracted from Wikidata.

4.3.1 Train/Test Splits

To build classifiers for the target attributes, we split data into training and testing. We made sure to include all instances of a specific entity exclusively in either training or testing data (based on unique QID), as to not have aliases or translations from the same entity present in both the training and testing data. Testing data was created by extracting only entities that have all normalized attributes available (i.e, without any missing value labeled as *XX* in any attribute). This resulted in having 80,130 complete entities for testing and the remaining 1,016,780 partially completed entries for training. Test data represents 7% of all available data.

5 Experiments and Results

We model the problem as a supervised classification task, and we train a Linear Supper Vector Machine (SVM) model for each of the attributes investigated. We then compare the Linear SVM evaluation results to the results of BERT, in which we look at two approaches. It is important to note that throughout the whole experiment, only the first name was used as input when training any of the gender classifiers, since any other names do not usually provide any indication of the gender of an individual. For evaluation, we use the precision, recall, and F-1 score as our metrics[12].

5.1 Linear SVM

To extract the appropriate features of the names, we specified an n-gram range of a min_length of 2, and max_length of 5 on the character level, to analyze the patterns in the groups of characters within each name, and identify certain patterns that could potentially influence the predictions of the classifier. As shown in table 3, there is a large imbalance in the number of male versus female instances in the data. Imbalance was also present in other attributes as well. As a result, we will report and reference the weighted F1-score throughout this paper.

We ran an additional experiment including deriving language and region information from the predicted country, as shown in Appendix E, to see whether this would yield better prediction results. This experiment showed that deriving language and region information from the SVM_Country

[12] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

Attributes	Available Entities		
	Original	Normalized	Balanced
Country	3.83M (41.28%)	3.04M (32.75%)	834K (8.99%)
Gender	7.34M (79.01%)	5.91M (63.61%)	821K (8.84%)
Language	1.92M (20.66%)	1.63M (17.60%)	410K (4.42%)
Region	NA	3.04M (32.75%)	834K (8.99%)

Table 2: Total number and percentage of the unique available entities (non-*XX*), in original, normalized, and balanced data for each investigated attribute, in relation to the unique original data extracted from Wikidata (9,284,039).

	Precision	Recall	F-1 score	No. of instances
female	0.95	0.72	0.81	8,755
male	0.97	0.99	0.98	71,375
accuracy			0.96	80,130
macro avg	0.64	0.57	0.60	80,130
weighted avg	0.96	0.96	0.96	80,130

Table 3: Detailed evaluation results for SVM_Gender classifier

prediction could not surpass the results of the SVM experiments. The derived_language predictions achieved an F-1 score of 0.72 (0.03 less than SVM_Language) while derived_region had an F-1 score of 0.64.

5.2 Transformers Approach

Transformers are well known for capturing information and semantics beyond the word-level. Such features motivate us to investigate how we could improve the scores further. We decided to experiment with two approaches related to Transformers.

5.2.1 Transformers Pipelines - Pretrained BERT Model

Using the Transformer package from huggingface.io, we approached this task as 'Masked Language Modeling (MLM) prediction'. We formulated the following text patterns that would potentially allow to predict the Country, Gender and Language of an individual from their name:[13]

- Country = unmasker("%s is from the country of [MASK]." % name)
- Gender = unmasker("%s was happy when [MASK] met me." % name)
- Language = unmasker("%s can speak the [MASK] language." % name)

It is worth mentioning that, in certain instances, BERT might generate non-standard predictions, and therefore, we had to manually revise all differences between expected results and predictions to include possible variations (ex: US vs USA vs America).

In addition, it is important to note that region was not included in this experiment since BERT predictions were hard to match with our predefined regions. As a result, we deduced the region information from the Country extracted. These predictions were made using the *bert-base-multilingual-uncased*[14] model.

After normalization of the results, as shown in table 5, we can see that the weighted F-1 scores for all attributes are lower than the scores reported in table 4. As a result, we decided to explore a different approach for this task by fine-tuning a BERT model.

5.2.2 BERT Model Fine-tuning

Due to the time required to fine-tune a BERT model, we decided to focus on the *Country* attribute since it is the attribute with the most classes, and has the most room for improvement when it comes to evaluation. We found that predicting the country when fine-tuning the *bert-base-multilingual-uncased* model, resulted in a weighted F-1 score of 0.54. Although this score is 0.01 higher than

[13] We tried different patterns and these are the best performing ones.

[14] https://huggingface.co/bert-base-multilingual-uncased

SVM Model	Number of Classes	Weighted F-1 Score
SVM_Country	100	0.53
SVM_Language	48	0.75
SVM_Region	18	0.65
SVM_Gender	2	0.96

Table 4: Weighted F-1 scores for each attribute's model, training a linear SVM on clean names

Masked Model	Weighted F-1 Score
Masked_Country	0.29
Masked_Language	0.63
Masked_Region	0.41
Masked_Gender	0.87

Table 5: Weighted F-1 scores for each attribute's model, masked language modelling prediction

the SVM model, we found the SVM model a more suitable option due to the prediction time which is a fraction of that is required for the transformer models.

5.3 Evaluating on Twitter Dataset

Typically, person names in Wikidata cover celebrities and famous persons. We wanted to test the performance of our models on names extracted from Twitter which include all sorts of diverse names that are not necessarily popular. To do so, we used FollowerWonk[15] which allows searching in user profiles based on their declared locations. We selected 15 countries covering different regions.[16] For each country, we randomly selected 100 person names from the 1,500 users provided by FollowerWonk.

The model SVM_Country achieved a weighted F-1 score of 0.54. This result confirms the accuracy and the validity of the results achieved on the Wikidata test set (Table 4).

5.4 Explainability

To understand the features that are taken into consideration when predicting the country of an individual's name, we use Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016). LIME is a method that explains the logic that classifiers use to predict classes, by assessing the weights of each word in the input and its contribution to the final prediction (Visani et al., 2020). For this, we focus on the SVM_Country model since it has the most number of classes, which could result in more variation in the results.

Figure 3 shows two names derived from the Wikidata test set and the Twitter dataset, and the LIME analysis that they underwent when given to the SVM_Country model. In this case, the darker the word is highlighted, the bigger its contribution to the final prediction. For instance, in the name *Jesse Zhang*, a Chinese name, the surname *Zhang* had the biggest contribution to the country classifier, which indeed predicted the country China. Such analysis while it confirms the results of the system used, it also provides more justification on how the decision was made.

6 Conclusion

In this paper we investigated the amount of information we can infer from an individual's name. This includes country, gender, language, and region information. We extracted information from a total of 9,284,039 entities from Wikidata, and normalized the data to include information from 100 countries, 48 languages, 18 regions, and 2 genders. After balancing the data, and preprocessing it appropriately, we ended up with a total of 1,096,910 Arabic and English entities. Using this data, we experimented with an SVM-based approach, using a Linear-SVM, as well as BERT-based approaches, and found while BERT-based results are very competitive, the SVM-based approach yielded the best results and efficiency. The SVM approach was also evaluated on an external dataset that consists of 1,500 names from 15 diverse countries extracted from Twitter, and yielded similar results, showing that the models are stable and generalizable.

We believe that there are many extensions to this work. Exploiting more languages on top of Arabic and English, which would improve the generalizability of the techniques we experimented with. Additional information recorded in Wikidata and using it in research can open the doors to different research and applications. Furthermore, it

[15] https://followerwonk.com/

[16] Countries are: DZ, AR, AU, BR, CA, CN, EG, FR, DE, IN, MX, RU, SA, ZA, and US.

Figure 3: LIME analysis for sample of names from Wikidata test set and Twitter dataset

is important to note that although a name could be connected with a certain country, that does not necessarily mean that they are ethnically from that country, which is especially true in ethnically diverse countries such as the United States. As a result, studying the ethnicity of a name, paired with the country they are from could be an interesting exploration. Other research related to modern migration and refugee communities are all topics that can be explored from such data.

7 Ethics and Social Impact

Due to the nature of this exploration and the data used, there are ethical and social implications, as well as biases that should be discussed.

7.1 Biases

Our models were trained using publicly available data from Wikidata, and usually, different regions and ethnic groups are not as well-represented in the data source as other groups. This means that less "popular" names might not be as common within the data, which could affect the prediction capabilities of the models once they are presented with such names. This could be due to a couple of factors including some countries having much smaller populations, and/or the digital divide between the regions, which could have an impact on the amount of information we have access to from those less "digitally-advanced" regions (Pick and Sarkar, 2016).

7.2 Potential Misuse

Knowing any individual's personal information could be used to promote discriminatory practices, in relation to nationality and gender specifically. In other words, potentially being able to know information such as the country, region and gender of an individual before even meeting them, could lead to some unfair treatment. For instance, according to a Belgium based study that surveyed Belgian Human Resources professionals, "Arab female applicant received lower job suitability ratings compared with equally qualified native/Belgian female and Maghreb/Arab male applicants when they applied for a high-cognitive demanding job" (Derous and Pepermans, 2019). In this case, our models could somewhat promote this practice, which would be a misuse of such technology.

However, it is important to note that many technologies could be considered "double-edged swords" in that they possess both positive and negative use-cases, depending on the intentions of the entities using them. Therefore, one could argue that such technology could be used for user-demographic research, as well as market and audience research, and being able to automate the process of "...Assigning demographically appropriate names to data-driven entities, such as personas, chat-bots, and virtual agents..." through using gender and country data (Jung et al., 2021). In other words, training such models, and providing such a data-set as the one detailed in this investigation,

could aid researchers in automating otherwise expensive and time-consuming processes.

References

Olatunji Abdulganiy, Moshood Issah, Noah Yusuf, A Muhammed, and Abdul-Rasheed Sulaiman. 2015. Personal name as a reality of everyday life: Naming dynamics in select african societies.

Rochana Chaturvedi and Sugat Chaturvedi. 2022. It's all in the name: A character based approach to infer religion. *ArXiv*, abs/2010.14479.

Eva Derous and Roland Pepermans. 2019. Gender discrimination in hiring: Intersectional effects with ethnicity and cognitive job demands. *Archives of Scientific Psychology*, 7:40–49.

Jerome Friedman. 2002. Stochastic gradient boosting. *Computational Statistics Data Analysis*, 38:367–378.

Saman Goudarzi. 2020. 'what's in a name? using first names as features for gender inference in twitter' by wendy liu derek ruths (2013). In *Identifying Gender and Sexuality of Data Subjects*. Https://cis.pubpub.org/pub/whats-in-a-name-using-first-names-as-features-for-gender-inference-in-twitter.

Bas Hofstra and Niek C de Schipper. 2018. Predicting ethnicity with first names in online social media networks. *Big Data & Society*, 5(1):2053951718761141.

Wenyi Huang, Ingmar Weber, and Sarah Vieweg. 2014. Inferring nationalities of twitter users and studying inter-national linking. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, page 237–242, New York, NY, USA. Association for Computing Machinery.

Soon-Gyo Jung, Joni Salminen, and Jim Jansen. 2021. All about the name: Assigning demographically appropriate names to data-driven entities.

Md Kowsher, Md Sanjid, Avishek Das, Mahid Ahmed, and Md Sarker. 2020. Machine learning and deep learning based information extraction from bangla names. *Procedia Computer Science*, 178:224–233.

Jinhyuk Lee, Hyunjae Kim, Miyoung Ko, Donghee Choi, Jaehoon Choi, and Jaewoo Kang. 2017. Name nationality classification with recurrent neural networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 2081–2087. AAAI Press.

Shervin Malmasi. 2014. A data-driven approach to studying given names and their gender and ethnicity associations. In *ALTA*.

Chibuike Mbarachi and Esther Igwenyi. 2018. Language, identity and the cultural context of names in selected nigerian novels. *International Journal of Language and Literature*, 6.

James Pick and Avijit Sarkar. 2016. Theories of the digital divide: Critical comparison. pages 3888–3897.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. New York, NY, USA. Association for Computing Machinery.

Cong Tang, Keith Ross, Nitesh Saxena, and Ruichuan Chen. 2011. What's in a name: A study of names, gender inference, and gender behavior in facebook. In *Database Systems for Adanced Applications*, pages 344–356, Berlin, Heidelberg. Springer Berlin Heidelberg.

Konstantinos Tzioumis. 2018. Demographic aspects of first names. *Scientific Data*, 5:180025.

Giorgio Visani, Enrico Bagli, and Federico Chesani. 2020. Optilime: Optimized lime explanations for diagnostic computer algorithms. *ArXiv*, abs/2006.05714.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

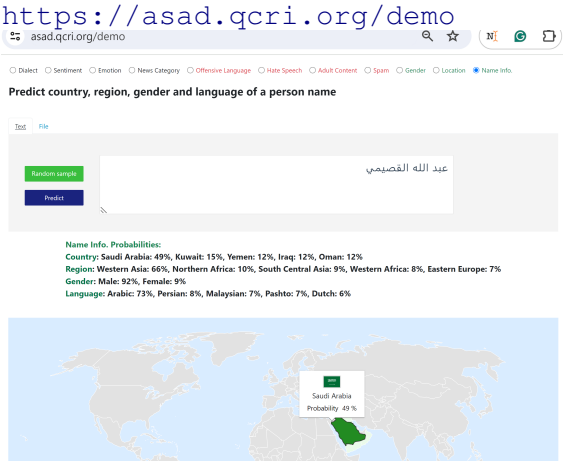
Stephen Wilson. 2003. The means of naming: A social and cultural history of personal naming in western europe. *The Means Of Naming: A Social And Cultural History Of Personal Naming In Western Europe*, pages 1–402.

Junting Ye, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, and Steven Skiena. 2017. Nationality classification using name embeddings. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1897–1906, New York, NY, USA. Association for Computing Machinery.

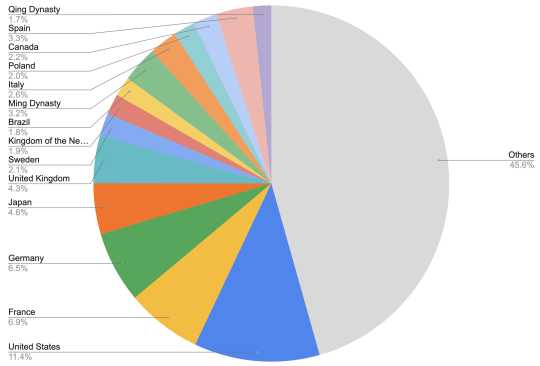
Junting Ye and Steven Skiena. 2019. The secret lives of names? name embeddings from social media. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '19, page 3000–3008, New York, NY, USA. Association for Computing Machinery.

A Demo of Functionality

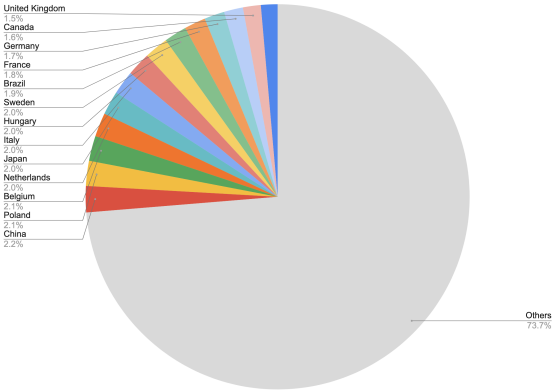
Demo Link (feature: NameInfo) - https://asad.qcri.org/demo



B Countries Distribution in Original Data Extracted From Wikidata.



C Countries Distribution in Balanced Data



D Gender Values

The top original gender values from Wikidata are listed here (35 values in total). Values other than "male", "female" and "XX" (empty) represent 0.03% of the data and we ignored them in our processing.

Original Value	Mapping
male	male
female	female
XX (empty)	- (ignored)
transgender female	-
non-binary	-
transgender male	-
eunuch	-
intersex	-
genderfluid	-
genderqueer	-

E Country, Region and Language Values

The first 10 countries, in alphabetical order, are shown in the following table:

Country	Region	Language
Afghanistan	South-central Asia	Pashto
Algeria	Northern Africa	Arabic
Angola	Middle Africa	Portuguese
Argentina	Latin America and the Caribbean	Spanish
Australia	Australia/New Zealand	English
Austria	Western Europe	German
Azerbaijan	Western Asia	Azerbaijani
Bahrain	Western Asia	Arabic
Bangladesh	South-central Asia	Bengali
Belarus	Eastern Europe	Belarusian