Nationality Classification using Name Embeddings

Junting Ye[1], Shuchu Han[1], Yifan Hu[2], Baris Coskun[3] *, Meizhu Liu[2], Hong Qin[1], Steven Skiena[1]
[1]Stony Brook University, [2]Yahoo! Research, [3]Amazon Web Services, New York, NY.
{juyye,shhan,qin,skiena}@cs.stonybrook.edu,{yifanhu,meizhu}@yahoo-inc.com

ABSTRACT

Nationality identification unlocks important demographic information, with many applications in biomedical and sociological research. Existing name-based nationality classifiers use name substrings as features and are trained on small, unrepresentative sets of labeled names, typically extracted from Wikipedia. As a result, these methods achieve limited performance and cannot support fine-grained classification.

We exploit the phenomena of homophily in communication patterns to learn *name embeddings*, a new representation that encodes gender, ethnicity, and nationality which is readily applicable to building classifiers and other systems. Through our analysis of 57M contact lists from a major Internet company, we are able to design a fine-grained nationality classifier covering 39 groups representing over 90% of the world population. In an evaluation against other published systems over 13 common classes, our F1 score (0.795) is substantial better than our closest competitor *Ethnea* (0.580). To the best of our knowledge, this is the most accurate, fine-grained nationality classifier available.

As a social media application, we apply our classifiers to the followers of major Twitter celebrities over six different domains. We demonstrate stark differences in the ethnicities of the followers of Trump and Obama, and in the sports and entertainments favored by different groups. Finally, we identify an anomalous political figure whose presumably inflated following appears largely incapable of reading the language he posts in.

ACM Reference format:
Junting Ye[1], Shuchu Han[1], Yifan Hu[2], Baris Coskun[3] [1], Meizhu Liu[2], Hong Qin[1], Steven Skiena[1] . 2016. Nationality Classification using Name Embeddings. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 10 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Nationality and ethnicity are important demographic categorizations of people, standing in as proxies to represent a range of cultural and historical experiences. Names are important markers of cultural diversity, and have often served as the basis of automatic nationality classification for biomedical and sociological research. For example, nationality from names has been used as a proxy to reflect genetic differences [5, 10] and public health disparity [6, 25] among groups. Nationality identification is also important in ads targeting, academic studies of political campaigns and social media analysis [3, 11]. Name analysis is often the only practical way to gather ethnicity/nationality annotations, because of privacy concerns.

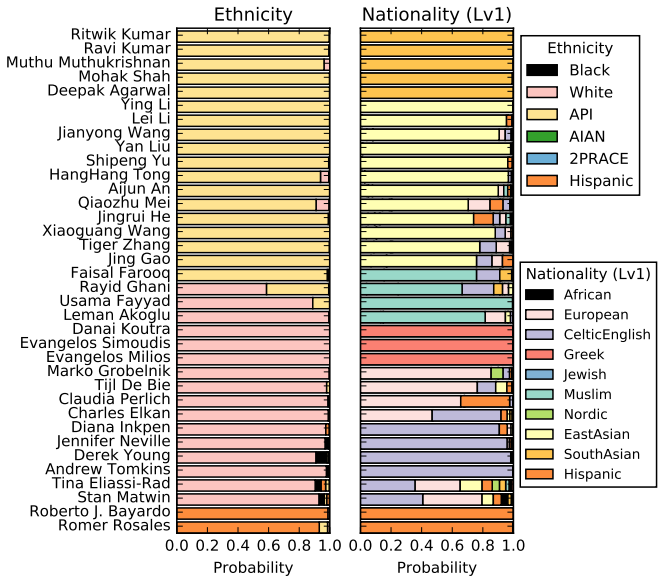

Figure 1: Ethnicity and nationality (Level 1 of the taxonomy) classification on some data mining researchers.

Several previous name-based ethnicity/nationality classification approaches have been presented [11, 28, 29], including [2] at KDD '09. However, the performance of these methods has been constrained by small and artifical training sets, such as celebrity names from Wikipedia, and restricted to coarse ethnicity/nationality taxonomies. The long tail of names makes these approaches dependent on surface forms (like substring distributions), which are by definition ineffective for logograms. Almost all existing methods are designed only for Latinized names, while other writing systems (e.g. Arabic, Cyrillic) are also widely used.

In this paper, we present *NamePrism*, a new name nationality and ethnicity classifier which offers a finer-grained taxonomy of ethnic groups. Fig. 1 demonstrates the performance of our system, by presenting the ethnicity/nationality probability distributions of some data mining researchers. We believe our results will generally agree with the reader's judgement.

Unlike previous methods that rely on substring features, we propose a more robust representation of names, which exploits the phenomenon of homophily in communication. The *homophily*

[1]This research was conducted when Baris Coskun was with Yahoo! Research.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Conference'17, Washington, DC, USA
© 2016 Copyright held by the owner/author(s). 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

principle, that people tend to associate with similar people or popularly that "birds of a feather flock together," is one of the most striking and empirically robust regularities in social life [15, 21]. Leskovec and Horvitz observed that, in instant messages, people tend to communicate more frequently with others of similar age, language and location [18]. We analyze over 57 million contact lists from an email company, where the account holders are anonymized. The homophily-induced coherence of these contact lists enables us to derive meaningful features using *word embedding* methods [22, 23] as the basis for a comprehensive and effective nationality classifier.

We collected 74M labeled names come from 118 different countries, containing over 90% of world's population. We use these labels to define a natural taxonomy of 39 leaf nationalities. As far as we know, our classifier is the most fine-grained and effective one accessible to the public. The main contributions of our work are:

- *Introducing Name Embeddings:* The contact-list derived name embeddings prove to be a powerful way to capture latent properties of gender, nationality, and age in features readily applicable to classification and regression tasks. Projections of these embeddings are very compelling, creating maps in embedding space that correspond to maps of national boundaries. We believe these embeddings will prove widely applicable to other applications and domains, including those in data privacy and security.
- *Improved Nationality Classification:* Our name-based nationality classifier *NamePrism* performs considerably better than previous classifiers. In particular, on a 13-class evaluation over email/Twitter data, our F1 score (0.795) proves to be much better than competing systems *Ethnea*[2] (0.580) [28], *HMM*[3] (0.364) [2], and (on a reduced 10-class scale) *EthnicSeer*[4] (0.571) [29]. *NamePrism* uses a Naive Bayes approach within a nationality taxonomy over 39 leaf nodes, employing name embeddings as the primary features.
- *Improved Ethnicity Classification:* A benefit of fine-grained nationality taxonomy is its flexibility to apply to different task settings.The six ethnic groups defined by U.S. Census Bureau over U.S. population largely corresponds to distinct nations of origin. Our ethnicity classifier *NamePrism*^{e}, simply reduces the nationality taxonomy from 39 leaf nodes to 6 and incorporates census-based ground truth parameters into the Naive Bayes model.
- *Online Classification Resources:* We release *NamePrism* as free web service[5] for research in sociology, linguistics, and biomedical applications. To the best of our knowledge, it is the only nationality classifier that handles various writing systems, and works on a fine-grained 39-class taxonomy.
- *Social Media Analysis:* We use *NamePrism* to analyze social media, specifically the followers' nationalities/ethnicities of 600 major celebrities on Twitter. Our results show that: (1) Donald Trump's U.S. followers are disproportionally White with followers of Obama and Clinton, (2) ethnicities exhibit different preferences in sports and entertainment, and (3) the follower counts of a particular Indonesian politician has been artificially inflated by Russian names.

The rest of this paper is organized as following. In Sec. 2, we introduce related works. Sect. 3 shows visualization and evaluations of name embeddings. In Sec. 4 and 5, we describe the methodology and experiments of *NamePrism* and *NamePrism*^{e}. We apply our methods on Twitter celebrities in Sec. 6.

2 RELATED WORK

Name nationality classification is a fundamental problem with a variety of important applications: *(i)* biomedical research and clinical practice: it is critical to study the genetic and dietary differences among distinct groups[5, 10]. *(ii)* sociology: health care/ employment/ education disparities among different people. [6, 25] *(iii)* online targeting: recommend more accurate ads/news/social media posts to users [3, 11]. Other applications includes population demographic studies [4, 16, 19, 20]. Despite wide-spread demand for nationality labels, it is hard to collect such information via self-reporting because of privacy concerns. Meanwhile, manual annotation of nationality by names is, in fact, a very difficult task, especially for fine-grained taxonomy.

Most recent works use name substrings as features for ethnicity/nationality classification [2, 11, 28, 29]. Ambekar et. al. [2] propose to combine decision tree and HMM to conduct classification on a taxonomy with 13 leaf classes. Treeratpituk et. al. [29] utilize both alphabet and phonetics sequences in names to improve performance and applied it to analyze how ethnicities evolves in computer science research community [31]. Chang et. al. [11] use Bayesian methods to infer ethnicity of Facebook users with U.S. census data and study the interactions between ethnic groups. Torvik and Agarwal [28] propose instance-based classifiers by using scientists' names from PubMed. In comparison, we propose name embedding in the light of homophily principle in social life [15, 18, 21]. It is a better representation because substrings are limited to phonogram. Other relevant efforts are binary ethnicity classifiers, including Hispanic [9], Chinese [12], South Asian [13].

Name embedding is inspired by word embedding[8, 22, 23], which has many applications in natural language processing [1, 7, 17]. Other types of data can also benefit from the same assumptions that underlie word embeddings, namely that a data point is governed by the other data in its context [24, 26, 27]. *DeepWalk* [24] learns node embeddings for graph data. It generates contexts by simulating random walks on graphs. Rudolph et. al. [26] propose a more general formulation of learning embeddings in different application settings. Similarly, name embeddings treats email contacts with most recency and frequency as context.

3 NAME EMBEDDINGS

Name embedding is a variation of word embedding. In a nutshell, word embedding algorithms [8, 22, 23] aim to learn similar embeddings (vectors) if two words co-occur frequently in their contexts. In articles, the context of a word are naturally the words around it. To generate context in contact lists, we need to assign orders to contacts. In the light of homophily principle, we weigh contacts by recency and frequency of communications. As a result, names with large weights tend to have same nationalities. In this way, we

[2]http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py
[3]http://www.textmap.com/ethnicity/
[4]http://singularity.ist.psu.edu/ethnicity
[5]**Open API**: http://www.name-prism.com/

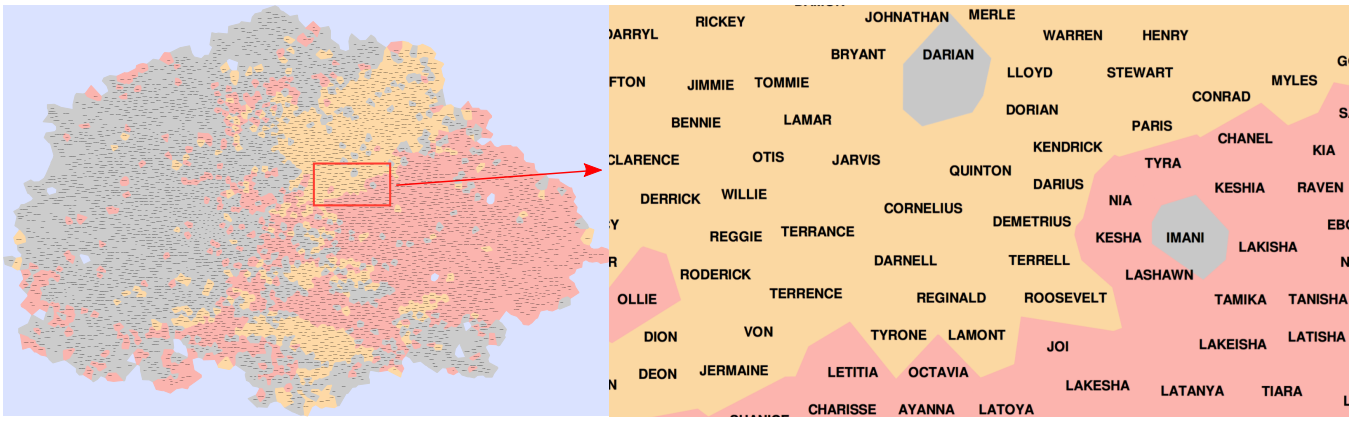

Figure 2: 2D projection (left) of 5K popular first names' *embeddings*. Orange are male names, salmon for females and gray for unlabeled. Same-gender names cluster together, indicating similar embeddings. Inset of the male-female border (right) shows more neutral names.

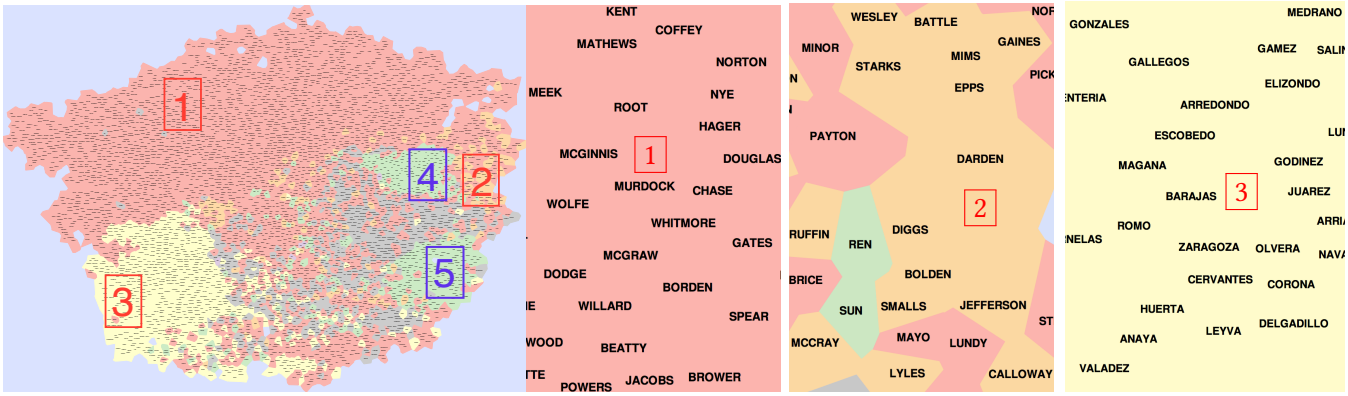

Figure 3: 2D projection (left) of 5K popular last names' *embeddings*. Same-ethnicity names stand close indicating similar embeddings. Insets (left to right) highlight *White* 1, *Black* 2 and *Hispanic* 3 names. *API* (4 and 5) in Fig. 4.

construct a "sentence" by keeping top contacts of a sorted list. Note that the ordering of sentences in an article is informative for word embeddings. In contrast, the ordering of the contact lists is not useful because email account holders are mutually independent.

3.1 Name Embedding Visualization

We use t-SNE [30] to project the 100D name embeddings into 2D, and create the map visualization with `gvmap` [14]. U.S. census data are used as ground truths to visualize and evaluate name embeddings. More specifically, we use U.S. 1990 Census data to label popular first names (4.7K female and 1.2K male) and U.S. 2000 Census data to label popular last names (115K White, 5K Black, 6K Asian/Pacific Islander (API), 0.2K American Indian/Alaskan Native (AIAN), 0.1K Two or more races (2PRACE) and 7K Hispanics). As shown from Fig. 2 to Fig. 4, genders and ethnicities are labeled with different colors. We are surprised to see how names with same gender, ethnicity and nationality cluster together (Fig. 2 to Fig. 4).

Fig. 2 (left) illustrates the landscape of first names. Using 1990 Census data, we color male names orange, female names pink, and names with unknown gender gray. In general, names of the same gender form mostly contiguous regions. Fig. 2 (right) is an inset showing a region along the male/female border. We can see that "Ollie" is labeled as a female name based on Census data (2:1 ratio of female/male instances), while in fact it is often used as a nickname for "Oliver" or "Olivia" for daily use. Therefore name embedding is correct in placing it near the border. The embedding also correctly placed "Imani" and "Darian", two names not labelled by the Census data, near the border, but in the female/male regions, respectively.

Fig. 3 (left) shows a map of last names. We color a name according to the dominant ethnicity classification from 2000 Census data. Four major ethnicities are White (pink), Black (orange), Hispanic (yellow), and API (green). Names beyond census data are colored gray. The three insets in Fig. 3 highlight the homogeneity of regions by ethnicities. White, Hispanic and API stand in large contiguous regions while Black are more dispersed. It makes sense because many Black people adopt White names during American slavery time. More interestingly, there are two distinct Asian regions in the map. Fig. 4 presents insets for these two regions, revealing that one

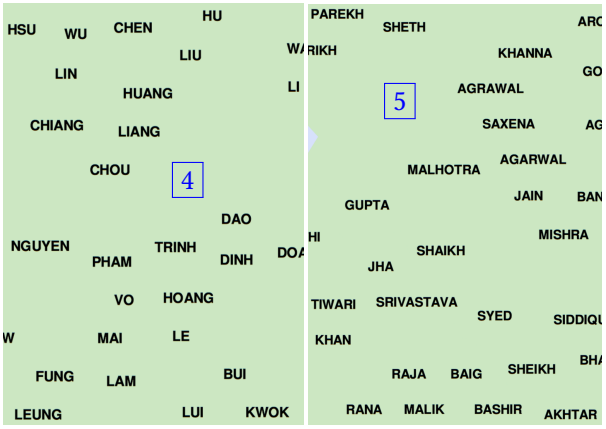

Figure 4: Two distinct Asian clusters. Left: Chinese/ Vietnamese names (4). Right: Indian names (5). It shows name embeddings capture nationality signals.

cluster consists of Chinese and Vietnamese names (left) while the other (right) contains Indian names. Even on the left subfigure, Vietnamese names are more gathering around the bottom part while Chinese names on the top. These observations strongly indicate name embeddings capture gender, ethnicity and nationality signals.

3.2 Evaluation

We run experiments to validate our observations quantitatively and explore the sensitivity of name embeddings under different parameters. The parameters that we test include: *(i)* different embedding learning method: *CBOW (Continuous Bag Of Word)* or *SG (Skip-Gram)*; *(ii)* use joint embedding space of first/last names or separate; *(iii)* number of nearest neighbor.

We can see from Tab. 1 that the joint variants generally perform best. However the differences between the variants are relatively small. In addition, the *CBOW* model generally outperforms the *SG* model. It seems $P_1(B|B)$ is relatively low (0.35-0.59). However, it is essentially a harder task to find a black name because a random name from the contact lists has a probability of 0.03 being Black, while 0.74 being White.

	Metrics	Joint		Seperate	
		CBOW	SG	CBOW	SG
Gender	$P_1(G_i\|G_i)$	0.909	0.884	**0.916**	0.884
	$P_{10}(G_i\|G_i)$	**0.936**	0.927	0.935	0.921
Ethnicity	$P_1(W\|W)$	0.936	**0.946**	0.930	0.922
	$P_1(B\|B)$	**0.594**	0.456	0.444	0.345
	$P_1(A\|A)$	**0.763**	0.721	0.717	0.680
	$P_1(H\|H)$	**0.754**	**0.754**	0.671	0.697

Table 1: Evaluations of different name embedding variants. CBOW and SG are two word embedding methods. $P_1(G_i|G_i)$ is the probability that 1 nearest neighbor (1-NN) is of the same gender while $P_{10}(G_i|G_i)$ is for 10-NN. "W", "B", "A", "H" stand for "White", "Black", "API", "Hispanic", respectively.

4 NATIONALITY CLASSIFICATION

4.1 Methodology

NamePrism uses Naive Bayes model because of its effectiveness and interpretability. We argue that name nationalities depend on both first name and last name. This is especially effective for names used across different nationalities but with different popularities. It also helps to reduce errors when names are mixtures because of immigration or cross-nationality marriages. We put much effort on estimating parameters, i.e. name parts likelihood, using features from training data, name embedding, substrings and string characters. Therefore, each parameter has at most 4 estimations. NamePrism uses the ones with largest confidence for predictions.

4.1.1 Naive Bayes Model. In many case, our last names reveal our nationality origins. For example, "Zhang" is a common Chinese last name. It is easy to predict one's nationality if his last name is unique to that nation. However, there are many last names that are popular across nationalities. For example, "Lee" is popular in both China (especially in Hong Kong) and the UK. For "Qiang Lee" and "John Lee", we would make mistakes if we only take signals from the last name. Combining with first names, we can perform better because it is easy to see whether the first name is more in China or UK. Similarly, using both name parts also helps when names are mixtures due to immigration or cross-nationality marriage.

Our method, *NamePrism*, can be formalized in Eq. 1:

$$P(N|v_f, v_l) \propto P(v_f|N)P(v_l|N)P(N) \tag{1}$$

where N denotes nationality, v_l means last name and v_f is first name. We will describe our methods to estimate the likelihood (i.e. $P(v_f|N)$, $P(v_l|N)$) for frequent and rare names in next subsection. We can get Equ. 1 by using Bayesian rule under the assumption that v_f and v_l are conditionally independent given N.

4.1.2 Parameter Estimation. We estimate name part likelihood from 4 sources: *(i) training data*, i.e. the names appear in training data (denoted as V_{tr}); *(ii) name embedding*, the names from contact lists that have embeddings (V_{em}); *(iii) prefix/suffix strings*, names that share the same prefix/suffix with names in training data ($V_{p/s}$); *(iv) name characters*, names that use the same language characters (e.g. Arabic) seen in training data (V_{ch}). Intuitively, the increasing order of vocabulary size is V_{tr}, V_{em}, $V_{p/s}$, V_{ch}, which is also the decreasing order of estimation confidence.

Training Data. Eq. 2 shows the most effective and simple way to estimate $P(v_l|N)$ and $P(v_f|N)$ directly from training data.

$$P_{tr}(v_i|N) = \frac{C(v_i, N)}{C(N)}, v_i \in V_{tr} \tag{2}$$

where v_i is either a first name or last name from V_{tr}. $C(v_i, N)$ is the count of v_i with nationality N and $C(N)$ is equivalent to $\sum_v C(v, N)$. Note that each name part in V_{tr} have more than 5 occurrences in training data so that we have high confidence in the estimation.

Name Embedding. The likelihood of names (v_i) in V_{em} can be estimated using k-NN, i.e. take the average of k nearest neighbors (e.g. kNNs) in V_{tr}. However, we did not directly estimate the likelihood using its kNNs' likelihood. Instead, we realize that it

performs better if we first estimate v_i's posterior using its neighbors' posteriors and then apply Bayes rule to estimate the likelihood (Eq. 3). It makes sense because names with similar embeddings do not necessarily have similar popularity (see Fig. 3). The estimation of $P_{em}(v_i|N)$ is formulated by Eq. 3 and 4.

$$P_{em}(v_i|N) = \frac{P_{em}(N|v_i)P(v_i)}{P(N)}, v_i \in V_{em} \tag{3}$$

$$P_{em}(N|v_i) = \frac{1}{|kNN(v_i)|} \sum_{v_j \in kNN(v_i)} P_{tr}(N|v_j) \tag{4}$$

where $kNN(v_i)$ is the set of name parts that are v_i' kNNs.

Prefix/Suffix Strings. As mentioned in [2], prefix and suffix of name parts are indicative features. For name part $v_i \in V_{p/s}$, we can estimate its likelihood by averaging the ones' which share the same prefix/suffix.

$$P_{p/s}(v_i|N) = \frac{1}{|PS(v_i)|} \sum_{v_j \in PS(v_i)} P_{tr}(v_j|N), v_i \in V_{p/s} \tag{5}$$

where $PS(v_i)$ is the set of prefix and suffix strings of v_i. Here we use substrings with length between 3 to 5. $P_{p/s}(v_j|N)$ is the average likelihood of name parts in V_{tr} that have prefix/suffix v_j.

Name Characters. If a name is so rare that it is not in V_{tr} nor V_{em}. Moreover, it doesn't contain valid prefix or suffix strings. For example, a name written in "Hangul", "근혜" . It is very likely to be a Korean name because most names in "Hangul" are Korean names. Therefore, for a name $v_i \in V_{ch}$, we use the average of names in same characters to estimate its likelihood.

$$P_{ch}(v_i|N) = \frac{1}{|CH(v_i)|} \sum_{v_j \in CH(v_i)} P_{tr}(v_j|N), v_i \in V_{ch} \tag{6}$$

where $CH(v_i)$ is the set of names in training data that are written in the same language as v_i.

4.1.3 Internet Population vs. World Population. As we have shown in previous subsections, the name parts likelihood are estimated from email/Twitter users. However, Internet services (Email and Twitter) has varying popularity in different countries. Therefore, we need to assign different priors if a name is not sampled from Internet users. For example, UK and South Africa have similar population (around 50M to 60M). In our datasets, we have an order of magnitudes more names from the UK than from South Africa. Therefore we need to adjust to the real population of countries when we are predicting a random name from the world population.

Formally, let $P^I(\cdot)$ be probabilities over Internet population, and $P^W(\cdot)$ be the probability over world population. We have $P^I(v_i|N) = P^W(v_i|N)$ by assuming that names of Internet population are random samples from corresponding countries. Let I^N be the number of names with nationality N on Internet population and W^N be the one of N on world population. Thus, $P^I(N) = \frac{I^N}{\sum_N I^N}$ and $P^W(N) = \frac{W^N}{\sum_N W^N}$. We can get the relation between $P^I(N)$ and $P^W(N)$ with Eq. 7.

Algorithm 1: *NamePrism* , a hierarchical nationality classifier

Input : first/last name v_f, v_l; nationality taxonomy; estimated parameter sets $P_{tr}, P_{em}, P_{p/s}, P_{ch}$.
Output: nationality prediction T

1 **Init** T = root class ;
2 **while** T *is not a leaf class* **do**
3 **for** *child class* N_i *of* T **do**
4 **for** *each name part* $v \in \{v_f, v_l\}$ **do**
5 **if** $v \in V_{tr}$ **then**
6 $P(v|N_i) = P_{tr}(v|N_i)$;
7 **else if** $v \in V_{em}$ **then**
8 $P(v|N_i) = P_{em}(v|N_i)$;
9 **else**
10 $P(v|N_i) = \sigma$; # σ is a small constant
11 **if** *neither of* v_f, v_l *in* V_{tr} *or* V_{em} **then**
12 **for** *child class* N_i *of* T **do**
13 **for** *each name part* $v \in \{v_f, v_l\}$ **do**
14 **if** $v \in V_{p/s}$ **then**
15 $P(v|N_i) = P_{p/s}(v|N_i)$;
16 **else if** $v \in V_{ch}$ **then**
17 $P(v|N_i) = P_{ch}(v|N_i)$;
18 $P(N_i|v_f, v_l) \propto P(v_f|N_i) \cdot P(v_l|N_i) \cdot P(N_i)$;
19 $T = \arg\max_{N_i} P(N_i|v_f, v_l)$;
20 **return** T;

$$P^I(N) = \frac{I^N}{\sum_N I^N} = \frac{W^N}{\sum_N W^N} \frac{\frac{I^N}{W^N}}{\frac{\sum_N I^N}{\sum_N W^N}} = P^W(N)\frac{s_N}{S} \tag{7}$$

where S is the overall sample ratio and s_N is the sample ratio of N. $P^I(N)$ can be estimated from training data. s_N and S can be computed by looking up countries' populations. We can put Eq. 7 into Eq. 1 when classifying names from world population.

4.1.4 Hierarchical Classification. Names are classified on a predefined taxonomy in top-down fashion (see Fig. 5). The detailed algorithm are shown in Alg. 1. We start from root class of the taxonomy (line 1). In each iteration, it picks the class that maximizes $P(N_i|v_f, v_l)$ (from line 2 to 19) until it meets a leaf class. Since we have higher confidence in P_{tr} than P_{em}, so we prefer parameters from the former (line 3 to 10). If neither of the name parts are in P_{tr} or P_{em}, we use the parameters from $P_{p/s}$ or P_{ch} (line 11 to 17). Note that if only one of the name part in P_{tr} or P_{em}, we will only use the partial signal and smooth the other name part.

4.2 Nationality Taxonomy Construction

The nationality taxonomy is a key component in our method. Mateos et al. proposed a nationality taxonomy based on Cultural, Ethnic and Linguist (CEL) similarities [20]. Our name-based nationality taxonomy is constructed on top of CEL-based taxonomy, especially for the top level construction. While there is no "gold

Figure 5: Treemap of nationality taxonomy. Nested blocks within a larger block are its child nodes. 118 countries/regions, covering over 90% world population, are assigned to 39 leaf nationalities. The taxonomy is constructed based on Cultural, Ethnic and Linguist (CEL) similarities.

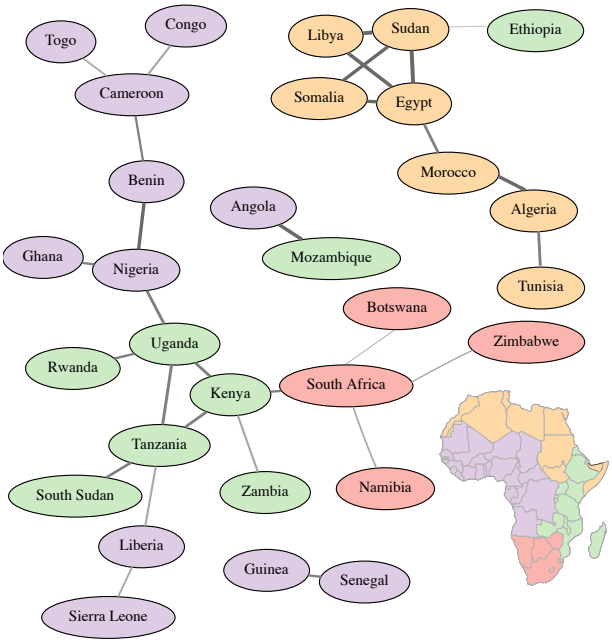

Figure 6: Name similarities between countries of Africa using Email/Twitter data. Thicker edges indicate stronger similarities and more common first/last names usages. Eastern African countries in green, western in purple, northern in orange and southern in red. The clusters of same-color nodes indicate that countries with similar names tend to be close geographically.

standard" name-based nationality taxonomy because of the complexity in naming customs around the world, we consult opinions from linguists and people from different cultures to reach a common ground as a useful approximation . Moreover, as shown in Sec. 4.3.2, we could compute similarities between countries using name parts distributions. These similarities are helpful to construct the bottom levels of the taxonomy. For example, Hispanic countries are divided into three subgroups: Spanish, Portuguese and Philippines. The reason is, countries within Spanish and Portuguese are very similar to each other according to name similarities, indicting finer-granularity groupings are not feasible and necessary.

4.3 Datasets

4.3.1 Name Labels. In order to estimate parameters mentioned above, we need name labels, i.e. full name and nationality pairs. We collected 68M such pairs from the Email source and 6M pairs from Twitter, totaling 74M labeled names from 118 major countries (Fig. 5). These countries take up over 90% of world population. To remove noise, we filter out names where both parts appear only once. 91% names remain. Note that we are interested in nationalities, thus immigration countries, including U.S., Canada and Australia, are not included in our dataset. To preserve privacy for the email data, the IDs of these users (e.g. email address) are removed. Furthermore, we only retain the counts of first/last names and countries. We used the full name and country labels solely for the purpose of performance measurement. They are not retained for classification.

Email. 90% of name labels come from email. Each name part appears at least twice so that typos and random strings are filtered. Note that the email contact lists and labeled names are different set of users. We set 5 as thresholds for both V_{tr} and V_{em}. It turns out $|V_{tr}|$ is 1.02M and $|V_{em}|$ is 4.09M. It makes sense because contact lists are names from many email companies and thus a larger population.

	Wikipedia Data						Email/Twitter Data					
Nationality	Name#	*HMM*	*Ethnea*	*Embd*	*Prism*	*Prism*^{w}	Name#	*HMM*	*Ethnea*	*Embd*	*Prism*	*Prism*^{w}
GreaterAfrican	11K	0.428	0.532	0.480	**0.543**	0.486	31K	0.269	0.389	0.554	**0.645**	0.622
GreaterEuropean	113K	0.863	0.903	0.927	**0.932**	0.899	225K	0.725	0.815	0.861	**0.920**	0.902
Asian	24K	0.654	0.670	0.711	0.745	**0.748**	123K	0.674	0.709	0.763	**0.910**	0.904
Muslim*	7K	0.380	0.563	0.538	**0.615**	0.611	13K	0.204	0.374	0.602	**0.612**	0.533
Africans*	4K	0.285	0.268	0.282	**0.314**	0.259	18K	0.174	0.288	0.458	0.636	**0.659**
WestEuropean	49K	0.631	0.724	0.709	0.747	**0.756**	143K	0.553	0.735	0.780	0.873	**0.878**
EastEuropean*	9K	0.488	0.517	0.466	0.575	**0.629**	38K	0.301	0.582	0.726	0.794	**0.812**
British*	44K	0.611	0.760	0.789	**0.794**	0.768	35K	0.361	0.578	0.627	0.648	**0.689**
Jewish*	11K	**0.313**	0.111	0.095	0.129	0.183	9K	0.097	0.361	0.301	**0.405**	0.387
GreaterEastAsian	15K	0.637	0.626	0.642	0.690	**0.706**	97K	0.625	0.656	0.713	**0.907**	0.895
IndianSubContinent*	9K	0.523	0.660	0.768	**0.769**	0.746	26K	0.438	0.721	0.855	**0.912**	0.903
Italian*	14K	0.521	0.543	0.595	**0.634**	0.613	11K	0.233	0.453	0.665	0.713	**0.763**
Hispanic*	11K	0.403	**0.600**	0.397	0.521	0.538	69K	0.432	0.724	0.676	0.850	**0.864**
Nordic*	5K	0.400	0.587	**0.713**	0.709	0.709	23K	0.303	0.653	0.767	**0.783**	**0.783**
French*	14K	0.428	0.523	0.602	0.600	**0.624**	27K	0.203	0.426	0.738	**0.769**	0.750
Germanic*	5K	0.254	0.410	0.401	0.403	**0.412**	13K	0.140	0.431	0.582	0.629	**0.653**
Japanese*	8K	0.646	**0.724**	0.456	0.547	0.695	57K	0.674	0.788	0.434	0.928	**0.939**
EastAsian*	7K	0.499	0.455	0.609	**0.621**	0.549	40K	0.270	0.340	0.723	**0.834**	0.811
Weighted Avg.	—	0.492	0.607	0.619	0.648	**0.651**	—	0.364	0.580	0.642	0.790	**0.795**

Table 2: F1 scores on a 13-leaf taxonomy. Existing methods: *HMM* [2] and *Ethnea* [28]; *Embd* only uses parameters from name embeddings; *Prism*^{w} is *NamePrism* with world population as priors. Nationalities on different levels of taxonomy are separated with bold lines. '*' marks leaf nationalities. *Weighted Avg.* is count-weighted average F1 of leaf nationalities.

Twitter. Although the email data offers the majority of name labels, its imbalanced popularity across the world make some regions inadequate name labels. We noticed that Twitter[6], as an emerging Web service, has a wider coverage and thus can act as a supplementary source of name labels.

In order to get name labels from interested regions, we *(i)* get list of most popular regional celebrities[7]; *(ii)* get all followers' Twitter profiles of the celebrities'. Each profile record contains "name" and "location" fields, though many users leave the latter blank. In summary, we gathered 43M unique Twitter user profiles, within which 9M have non-empty "location" field and well-formed names (e.g. two name parts and string length > 1). However, these location tags are not well defined. Among 9M profiles, there are ~1.5M unique locations. Some of them are simply noise, while some offer too much details (e.g. university name without country info.). Therefore, we use Google Map API[8] to query for country names using the 10% most popular "locations". As a result, we have ~6M labeled names for use, supplementary to the labels from email source.

4.3.2 Name Similarities between countries. Since our labeled names are collected from Internet, it is important to check its quality. In this subsection, we provide an interesting perspective to validate the high quality of the datasets.

We compute the similarities between countries using the aggregations of names, and check whether they agree with common sense. In fact, we observe that the cultural/spatial closeness between countries are well captured by country name similarities. Take African continent as an example (shown in Fig. 6). On the right-bottom part of the figure, the continent map is divided into 4 major parts based on how close they are culturally and geographically. On the remaining part of the figure, countries with names labels are colored in accordance with the map. It is apparent that countries with same colors are clustered, indicating that nearby countries have similar names. One interesting case is that Angola is connected with Mozambique, even though one is on the west coast of the continent while the other is on the east coast. The reason is that both countries were once colonized by Portuguese, thus many Internet users have Portuguese names.

We compute the similarities between countries with following steps: (i) aggregate name parts of each country so that countries are represented by name part vectors, where each dimension indicates how many name parts occur in the countries, (ii) compute cosine similarity between vectors, i.e. name similarities between countries. Note that in Fig. 6, the thickness of edges indicate the magnitude of similarities. One link is made if either the similarity is larger than 0.5 or it makes sure that each country is linked to at least one most similar countries. Therefore, Ethiopia is linked to Sudan with a very small weight, though it is distinct from other countries.

4.4 Performance Evaluation

In this Subsection, we will first compare our method with existing systems on smaller nationality taxonomies (one 13-leaf taxonomy and one 10-leaf flat taxonomy [2, 28, 29]). Note we use their Web APIs to collect the classification results. Two independent datasets are tested on. The smaller one is from Wikipedia (used in [2, 29]), the other is from our test set of labeled names. In the end, we

[6]Twitter API: https://dev.twitter.com/rest/public
[7]https://www.socialbakers.com/statistics/twitter/profiles/kenya/
[8]https://developers.google.com/maps/

	Wikipedia			Email/Twitter		
Nationality	Name#	*Seer*	*Prism*	Name#	*Seer*	*Prism*
Muslim	7K	0.560	**0.646**	13K	0.422	**0.688**
EastEuropean	9K	**0.739**	0.596	38K	0.343	**0.804**
British	44K	**0.852**	0.843	35K	0.577	**0.726**
Indian	9K	0.768	**0.779**	26K	0.639	**0.880**
Hispanic	11K	**0.605**	0.558	69K	0.610	**0.871**
Germanic	5K	0.464	**0.487**	13K	0.433	**0.694**
French	14K	**0.676**	0.650	27K	0.482	**0.802**
Italian	14K	**0.707**	0.641	11K	0.329	**0.728**
EastAsian	7K	**0.824**	0.635	40K	0.418	**0.848**
Japanese	8K	**0.875**	0.550	57K	0.902	**0.929**
Weighted Avg.	—	**0.751**	0.700	—	0.571	**0.831**

Table 3: F1 scores on 10 nationalities. *EthnicSeer*[29] performs slightly better on Wikipedia data but it is an unfair comparison because it is trained on the same dataset. *NamePrism* performs significantly better on a larger test set from Email/Twitter.

will introduce more details about *NamePrism*'s performance on a finer-grained nationality taxonomy.

4.4.1 On Small Taxonomy. Ambekar et al. proposed an HMM-based method, which used signals from substrings of names [2] to classify name nationalities. Their taxonomy contains 13 leaf nodes and 18 nodes in total (see [2] for the definition of this taxonomy). In order to compare, all methods need to be on the same taxonomy. HMM is designed on this taxonomy. *NamePrism* and *Ethnea* are adapted to this because both methods are defined on a finer-grained taxonomy. *EthnicSeer* is compared separately on a flat 10-nationality taxonomy.

Two datasets are available for comparison: *(i)* the labeled names from Wikipedia (150K in total, the same dataset used to train *HMM* and *EthnicSeer*); *(ii)* we divide Email/Twitter data into training and testing datasets (60% vs. 40%). Then we sample 2% from the test data for use because it is not efficient to get classification results of baselines from their Web APIs (380K). Some small nationalities are given larger sampling ratio to get large enough test samples.

As shown in Tab. 2, we compare results of five methods: *HMM* [2], *Ethnea* [28], *Embd*, *NamePrism* and *NamePrism*^{w}. *Embd* only use parameters estimated from name embeddings. *NamePrism*^{w} uses the world population as priors. *NamePrism* and *NamePrism*^{w} performs best on most classes for both datasets. On Wikipedia data, our methods achieves best performances on 15 (out of 18) classes. Some classes get +10% F1 boost, including Indian, Nordic and East-Asian. On Email/Twitter data, the improvement is more significant. *NamePrism* outperforms the rest on all classes. Some classes get performance increase by +30%, including Muslim, Africans, etc. Note that *Embd* also achieves considerable high performance, indicating that name embedding is capturing nationality signals well.

EthnicSeer is defined on a 10-leaf flat taxonomy. For comparison purpose, we removed the labeled names from African, Jewish and Nordic from both datasets. We also shrink *NamePrism*'s 39-leaf taxonomy to fit this small one. The weighted average F1 score shows *EthnicSeer* performs slightly better on Wikipedia but it is the same dataset that *EthnicSeer* is trained on. In contrast, *NamePrism* performs significantly better on Email/Twitter testing set.

Nationality	Name#	*Prism*	Nationality	Name#	*Prism*
CelticEnglish*	3505K	0.725	SouthAsian*	2623K	0.890
Jewish*	11K	0.396	African	606K	0.589
Muslim	1475K	0.741	EastAsian	6157K	0.920
Greek*	259K	0.887	Hispanic	6892K	0.907
Nordic	195K	0.731	European	5371K	0.836
Nubian*	577K	0.650	Japan*	65K	0.836
Maghreb*	47K	0.148	Malay	2596K	0.863
ArabPeninsula*	172K	0.510	Chinese*	2901K	0.928
Turkic	78K	0.676	Portuguese*	2683K	0.886
Pakistanis	179K	0.511	Philippines*	1137K	0.724
Persian*	423K	0.656	Spanish*	3072K	0.851
Finland*	30K	0.739	German*	1278K	0.739
Scandinavian	165K	0.704	Baltics*	12K	0.408
WestAfrican*	315K	0.563	French*	2674K	0.825
SouthAfrican*	66K	0.370	Russian*	121K	0.716
EastAfrican*	225K	0.574	EastEurope*	65K	0.492
SouthKorea*	68K	0.861	SouthSlavs*	68K	0.570
Indochina	528K	0.901	Italian	1153K	0.745
CentralAsian*	3K	0.196	Cambodia*	1K	0.162
Turkey*	75K	0.687	Vietnam*	502K	0.913
Bangladesh*	78K	0.578	Thailand*	18K	0.592
Pakistan*	101K	0.449	Malaysia*	242K	0.480
Denmark*	49K	0.662	Indonesia*	2354K	0.870
Sweden*	74K	0.607	Romania*	329K	0.663
Norway*	42K	0.620	Italy*	825K	0.710
Myanmar*	7K	0.607			
Weighted Avg.	—	0.806			

Table 4: *NamePrism* performance (F1 scores) on a 39-leaf nationality taxonomy. Nationalities in different levels are separated with bolder lines. '*' marks leaf nationalities. *Weighted Avg.* is count-weighted average F1 of leaf nationalities.

4.4.2 On Large Taxonomy. Tab. 4 shows *NamePrism* F1 scores on the large nationality taxonomy. Note we randomly split the Email/Twitter data into training and testing sets (60% vs. 40%) for 3 times. All reported performances of our methods (i.e. *Embd*, *NamePrism* and *NamePrism*^{w}) are average F1 of 3 runs. The standard deviations are all below 0.005. As we can see from Tab. 4, *NamePrism* performs well on most nationalities. For some less developed countries with few Internet users, including Central Asian countries and Maghreb countries, we have limited number of name labels and contact lists. Thus the performances on these nationalities are limited. To the best of our knowledge, our work is the first effort trying to classify names belonging to these regions.

5 ETHNICITY CLASSIFICATION

As we have mentioned in Sec. 3, U.S. Census Bureau defined 6 race/ethnicity: White, Black, API, Hispanic, AIAN and 2PRACE. In order to build classifier for these ethnicities, we need labeled names for these ethnicities to estimate parameters. Fortunately, U.S. Census Bureau published ethnicity distribution for popular

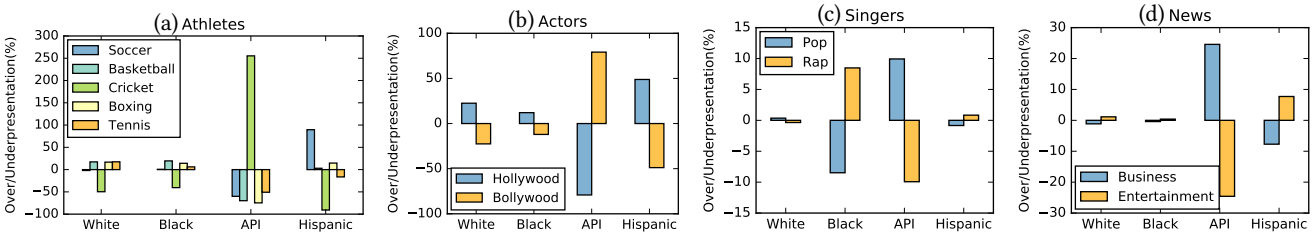

Figure 7: Ethnicity Over/underrepresentation of U.S. Twitter users' interest on different topics: (a) Cricket is almost exclusively followed by Indians while soccer is more popular among Hispanics. (b) U.S. actors enjoy a more diverse popularity than Indian actors. (c) African-Americans like rap more than pop. (d) Asians follow business news more than entertainment.

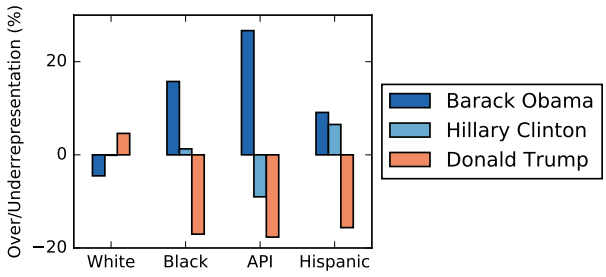

Figure 8: Ethnicity Over/underrepresentation of Barack Obama, Hillary Clinton and Donald Trump's U.S. Twitter followers. White followers are overrepresented for Trump, Obama and Clinton have more followers among minorities.

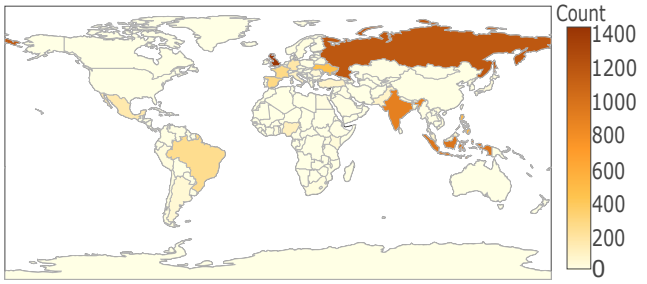

Figure 9: An anomaly Indonesian politician has surprisingly global impact: 50% of his followers have British, Russian or Indian names while only 13% are Indonesians. To validate, we find his Twitter profile suspicious: 23K Tweets since 2012 but only 1 following. His tweets are written in Indonesian, which most followers can not understand.

last names. We can estimate first names' ethnicity distribution by connecting census labels with email names from the U.S.

More formally, let V_{us}^{L} be the set of popular last names from Census Bureau, so we have ground truth, $P_{us}(E|v_l), \forall v_l \in V_{us}^{L}$, where E denote ethnicity. We can estimate the posteriors of first names with Eq. 8.

$$P_{us}(E|v_f) = \frac{1}{|S(v_f)|} \sum_{v_l \in S(v_f)} P(E|v_l) \tag{8}$$

where $S(v_f)$ is the list of last names and (v_f, v_l) is a full name from U.S. email data. Note that some of the last names paired with v_f may not have a ground truth label (i.e. $v_l \notin V_{us}^{L}$). To make reliable estimation, we only keep first names that at least half of the paired last names with a ground truth label. Therefore, we form a set of first names (V_{us}^{F}) with estimated ethnicity distributions. We denote $V_{us} = V_{us}^{F} \cup V_{us}^{L}$. We can get $P_{us}(v_f|E)$ and $P_{us}(v_l|E)$ by applying Bayes Rules.

For now, V_{us} can handle names with popular first/last names. For rare names, we can make use of the Email/Twitter name labels. 118 countries are assigned to the six ethnicities based on their definitions. For example, we make names from European countries as White while names from Asian as API. Therefore, we can follow similar steps as Algorithm 1. The difference is we will first check whether a name part is from V_{us}. If yes, we will use $P_{us}(v_i|E)$ to compute $P(E|v_f, v_l)$ because they are estimated from ground truth with high confidence. Otherwise, we will then check whether they are in V_{tr} or V_{em} as in Algorithm 1 and follow the remaining steps.

6 SOCIAL MEDIA ANALYSIS

Nationality and ethnicity classification have broad application in sociological research and media analysis. Here we present some interesting observations, when we apply our classifiers to the followers of Twitter celebrities.

To collect data, we identified the 100 most followed celebrities in each of six categories: actors, singers, news, atheletes, governments and politicians; all of whom have from 1M to 100M followers. For each celebrity, we selected 50,000 random followers, and filtered out accounts with irregular names using the same method as discussed in 4.3.1). We then apply *NamePrism* and *NamePrism*^{e} to the remaining followers. Our primary observations here include:

- *Ethnicity and the 2016 U.S. Presidential Election* – There has been considerable concern that the recent election exacerbated tensions between ethnic groups in the United States. Indeed, our analysis of U.S.-based followers of the primary figures in the race (Obama, Clinton, and Trump) show stark differences in composition. Fig. 8 shows that whites are substantially overrepresented among Trump's followers, while Clinton and Obama have disproportionately more followers among minorities.

- *Interests and Ethnicity* – Fig. 7 similarly breaks down the followers of major celebrities in sports, entertainment, and news categories. The followers of cricket and Bollywood stars are overwhelmingly Indian, while Hispanics disproportionally favor soccer and boxing.
- *Anomaly Detection through Nationality Analysis* – We were surprised to learn that an Indonesian politician named Jeffrie Geovanie was one of the most heavily followed figures on Twitter, because he has only 45K Google search results about him, mostly in Indonesian. Yet our name analysis of his followers shows that only 13% are Indonesian, with over 50% of the followers of British, Russian, or Indian nationality. This is quite peculiar given that Indonesian is the primary language of his Twitter stream.

7 CONCLUSION

We demonstrate that homophily patterns in communications can be exploited to learn name embeddings, that capture interesting properties of gender, nationality and ethnicity. Further we use these embeddings to build state-of-the-art name nationality and ethnicity classifiers. Through extensive experiments, we show that *NamePrism* substantially outperforms exiting methods on two independent datasets. Finally, we apply our classification to the Twitter celebrities' followers, with interesting results.

We believe that *NamePrism* will become an important tool for biomedical and sociological research. Future work revolves around applying name embeddings to other classification tasks, such as those arise in demographics, security and social media analysis.

REFERENCES

[1] Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. *ACL* (2013).

[2] Anurag Ambekar, Charles Ward, Jahangir Mohammed, Swapna Male, and Steven Skiena. 2009. Name-ethnicity classification from open sources. In *SIGKDD*. ACM, 49–58.

[3] Osei Appiah. 2001. Ethnic identification on adolescents' evaluations of advertisements. *Journal of Advertising Research* 41, 5 (2001), 7–22.

[4] Elizabeth Aries and Kimberly Moorehead. 1989. The importance of ethnicity in the development of identity of Black adolescents. *Psychological Reports* 65, 1 (1989), 75–82.

[5] Yambazi Banda, Mark N Kvale, Thomas J Hoffmann, Stephanie E Hesselson, Dilrini Ranatunga, Hua Tang, Chiara Sabatti, Lisa A Croen, Brad P Dispensa, Mary Henderson, et al. 2015. Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. *Genetics* 200, 4 (2015), 1285–1295.

[6] Donald A Barr. 2014. *Health disparities in the United States: Social class, race, ethnicity, and health.* JHU Press.

[7] Yoshua Bengio and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. *ICML* (2015).

[8] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3, Feb (2003), 1137–1155.

[9] Robert W Buechley. 1976. Generally useful ethnic search system: GUESS. In *Annual Meeting of the American Names Society*.

[10] Esteban González Burchard, Elad Ziv, Eliseo J Pérez-Stable, and Dean Sheppard. 2003. The importance of race and ethnic background in biomedical research and clinical practice. *The New England journal of medicine* 348, 12 (2003), 1170.

[11] Jonathan Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. 2010. ePluribus: Ethnicity on Social Networks. *ICWSM* 10, 18–25.

[12] Andrew J Coldman, Terry Braun, and Richard P Gallagher. 1988. The classification of ethnic status using name information. *Journal of epidemiology and community health* 42, 4 (1988), 390–395.

[13] Seeromanie Harding, Howard Dews, and Stephen Ludi Simpson. 1999. The potential to identify South Asians using a computerised algorithm to classify names. *Population Trends London* (1999), 46–49.

[14] Yifan Hu, Emden Gansner, and Stephen Kobourov. 2010. Visualizing graphs and clusters as maps. *IEEE Computer Graphics and Applications* 30 (2010), 54–66.

[15] Gueorgi Kossinets and Duncan J Watts. 2009. Origins of homophily in an evolving social network 1. *American journal of sociology* 115, 2 (2009), 405–450.

[16] Diane S Lauderdale and Bert Kestenbaum. 2000. Asian American ethnic identification by surname. *Population Research and Policy Review* 19, 3 (2000), 283–300.

[17] Quoc V Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents.. In *ICML*, Vol. 14.

[18] Jure Leskovec and Eric Horvitz. 2008. Planetary-scale views on a large instant-messaging network. In *WWW*. ACM, 915–924.

[19] Pablo Mateos. 2007. A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place* 13, 4 (2007), 243–263.

[20] Pablo Mateos, Richard Webber, and PA Longley. 2007. The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names. (2007).

[21] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*. 3111–3119.

[23] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–43.

[24] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *SIGKDD*. ACM, 701–710.

[25] James Quesada, Laurie Kain Hart, and Philippe Bourgois. 2011. Structural vulnerability and health: Latino migrant laborers in the United States. *Medical Anthropology* 30, 4 (2011), 339–362.

[26] Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. 2016. Exponential Family Embeddings. In *NIPS*. 478–486.

[27] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1067–1077.

[28] Vetle I Torvik and Sneha Agarwal. 2016. Ethnea–an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database. *International Symposium on Science of Science* (2016).

[29] Pucktada Treeratpituk and C Lee Giles. 2012. Name-ethnicity classification and ethnicity-sensitive name matching.. In *AAAI*.

[30] Laurens Van Der Maaten. 2014. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research* 15, 1 (2014), 3221–3245.

[31] Zhaohui Wu, Dayu Yuan, Pucktada Treeratpituk, and C Lee Giles. 2014. Science and Ethnicity: How Ethnicities Shape the Evolution of Computer Science Research Community. *arXiv preprint arXiv:1411.1129* (2014).