Assessing Demographic Bias in Named Entity Recognition

Shubhanshu Mishra*
smishra@twitter.com
Twitter, Inc.

Sijun He*
she@twitter.com
Twitter, Inc.

Luca Belli*
lbelli@twitter.com
Twitter, Inc.

ABSTRACT

Named Entity Recognition (NER) is often the first step towards automated Knowledge Base (KB) generation from raw text. In this work, we assess the bias in various Named Entity Recognition (NER) systems for English across different demographic groups with synthetically generated corpora. Our analysis reveals that models perform better at identifying names from specific demographic groups across two datasets. We also identify that debiased embeddings do not help in resolving this issue. Finally, we observe that character-based contextualized word representation models such as ELMo results in the least bias across demographics. Our work can shed light on potential biases in automated KB generation due to systematic exclusion of named entities belonging to certain demographics.

CCS CONCEPTS

• **Information systems** → *Computing platforms*; • **Computing methodologies** → **Information extraction**; • **Social and professional topics** → **Race and ethnicity**; **Gender**.

KEYWORDS

Datasets, Natural Language Processing, Named Entity Recognition, Bias detection, Information Extraction

ACM Reference Format:
Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing Demographic Bias in Named Entity Recognition. In *Proceedings of the Bias in Automatic Knowledge Graph Construction - A Workshop at AKBC 2020, June 24, 2020, Virtual.* ACM, New York, NY, USA, 12 pages.

1 INTRODUCTION

In recent times, there has been growing interest around bias in algorithmic decision making and machine learning systems, especially on how automated decisions are affecting different segments of the population and can amplify or exacerbate existing biases in society [18]. While many of the NLP ethics research papers focus on understanding and mitigating the bias present in embeddings [3, 7], bias in Named Entity Recognition (NER) [15, 16, 27] is not scrutinized in the same way. NER is widely-used as the first step of a variety of NLP applications, ranging from large-scale search systems [21] to automated knowledge graphs (KG) and knowledge base (KB) generation [9]. Bias in the first step of a pipeline could propagate throughout the entire system, leading to allocation and representation harm [1].

While most prior work focused on bias in embeddings, previous work has not given much attention to bias in NER systems. Understanding bias in NER systems is essential as these systems are used for several downstream NLP applications. To fill this gap, we analyze the bias in commonly used NER systems.

In this work, we analyze widely-used NER models to identify demographic bias when performing NER. We seek to answer the following question: *Other things held constant, are names commonly associated with certain demographic categories like genders or ethnicities more likely to be recognized?*

Our contributions in this paper are the following:

(1) Propose a novel framework[1] to analyze bias in NER systems, including a methodology for creating a synthetic dataset using a small seed list of names.
(2) Show that there exists systematic bias of existing NER methods in failing to identify named entities from certain demographics.

2 EXPERIMENTAL SETUP

Our general experimental setup is based on using synthetically generated data to assess the bias in common NER models, which includes popular NER model architectures trained on standard datasets and off-the-shelf models from commonly-used NLP libraries. As discussed in Section 2.1, we create the dataset with controlled context so that the effect of the names are properly marginalized and measured. We perform inference with various models on the dataset to extract person named entities and measure the respective accuracy and confidence of the correctly extracted names. Since capitalization is considered as an important feature for NER, we repeat the experiment with and without the capitalization of the name.

2.1 Data Generation and Pre-processing

In order to assess the bias in NER across different demographic groups, we need a corpus of sentences in which the named entity is equally likely to be from either demographic category. We overcome this issue by using sentence templates with placeholders to be filled with different names. In this work we only focus on **unigram person named entities**. Below we outline our approach for generating named entity corpora from two types of sentence templates. Using the same sentence with different names allows us to remove the confounding effect introduced by the sentence structure.

Names. Our name collection consists of 123 names across 8 different demographic groups, which are a combination of race[2] (or ethnicity) and gender. The categories span racial (or ethnic) categories, namely, Black, White, Hispanic, and Muslim [3]. For each race we include two gender categories, namely, male and female. Each demographic category, is represented in our name collection

*All authors contributed equally to this research.

[1] Details will be available at: https://github.com/napsternxg/NER_bias
[2] https://www.census.gov/topics/population/race/about.html
[3] We include Muslim and Hispanic along with other racial categories to better organize our results. We are aware that they are not racial categories.

KG-Bias '20, June 24, 2020, Virtual
.

category	Names
Black Female (BF)	Aaliyah, Ebony, Jasmine, Lakisha, Latisha, Latoya, Malika, Nichelle, Nishelle, Shanice, Shaniqua, Shereen, Tanisha, Tia, Yolanda, Yvette
Black Male (BM)	Alonzo, Alphonse, Darnell, Deion, Jamel, Jerome, Lamar, Lamont, Leroy, Lionel, Malik, Terrence, Theo, Torrance, Tyree
Hispanic Female (HF)	Ana, Camila, Elena, Isabella, Juana, Luciana, Luisa, Maria, Mariana, Martina, Sofia, Valentina, Valeria, Victoria, Ximena
Hispanic Male (HM)	Alejandro, Daniel, Diego, Jorge, Jose, Juan, Luis, Mateo, Matias, Miguel, Nicolas, Samuel, Santiago, Sebastian, Tomas
Muslim Female (MF)	Alya, Ayesha, Fatima, Jana, Lian, Malak, Mariam, Maryam, Nour, Salma, Sana, Shaista, Zahra, Zara, Zoya
Muslim Male (MM)	Abdullah, Ahmad, Ahmed, Ali, Ayaan, Hamza, Mohammed, Omar, Rayyan, Rishaan, Samar, Syed, Yasin, Youssef, Zikri
White Female (WF)	Amanda, Betsy, Colleen, Courtney, Ellen, Emily, Heather, Katie, Kristin, Lauren, Megan, Melanie, Nancy, Rachel, Stephanie
White Male (WM)	Adam, Alan, Andrew, Brad, Frank, Greg, Harry, Jack, Josh, Justin, Matthew, Paul, Roger, Ryan, Stephen
OOV Name	**Syedtiastephen**

Table 1: Name lists from different demographics.

with 15 salient names (and one with 16 names). A detailed list of names and their demographic categories is provided in Table 1.

Our name collection is constructed from two different sources. The first source of names comes from popular male and female first names among White and Black communities and was used to study the effect of gender bias in resume reviews in the work by Bertrand and Mullainathan [2]. This name dataset was constructed based on the most salient names for each demographic groups among the baby births registered in Massachusetts between 1974 and 1979[4]. The second source contains names in all eight demographic categories and is taken from the ConceptNet project[5] [25]. This collection of names was used to debias the ConceptNet embeddings [26]. We introduce a baseline name category to measure the context-only performance of the NER models with uninformative embedding. As described later, we also trained a few models in-house, for those models we directly use the OOV token. For pre-trained models, we use **Syedtiastephen**, which is unlikely to be found in the vocabulary but has the word shape features of a name. Hispanic names were deaccented (i.e. *José* becomes *Jose*) because including the accented names resulted in a higher OOV rate for Hispanic names.

We are aware that our work is limited by the availability of names from various demographics and we acknowledge that individuals will not-necessarily identity themselves with the demographics attached to their first name, as done in this work. Furthermore, we do not endorse using this name list for inferring any demographic attributes for an individual because the demographic attributes are personal identifiers and this method is error prone when done at an individual level. For the sake of brevity unless explicitly specified, we refer to names in our list by the community they are most likely to be found as specified in table 1. This means that when we refer to **White Female Names** we mean names categorized as **White Female** in table 1.

Among our name collections, the names **Nishelle (BF)**, **Rishaan (MM)**, and **Zikri (MM)** are not found in the Stanford GloVe [19] embeddings' vocabulary. Furthermore, the names **Ayaan (MM)**, **Lakisha (BF)**, **Latisha (BF)**, **Nichelle (BF)**, **Nishelle (BF)**, **Rishaan (MM)**, and **Shereen (BF)** are not found in the ConceptNet embedding vocabulary.

Winogender. Now we describe how to on generate synthetic sentences using sentence templates. We propose to generate synthetic sentences using the sentences provided by the Winogender Schemas [23] project. The original goal of Winogender Schemas is to find gender bias in automated co-reference solutions. We modify their templates to make them more appropriate for generating synthetic templates using named entities. Our modification included removing the word *the* before the placeholder in the templates and removing templates which have less than 3 placeholders. Examples of the cleaned up template and samples generated by us is shown in Table 2. We generated samples by replacing instance of *$OCCUPATION*, *$PARTICIPANT* and *$NOM_PRONOUN* in the templates with the names in our list, thus stretching their original intent. This gives us syntactically and semantically correct sentences. We utilize all triples of names, for each sentence template resulting in a corpus of $3! * \binom{123}{3} = 217$ million unique sentences.

In-Situ. To investigate the performance of the models on names in real world (or in-situ) data, we synthesize a more realistic dataset by performing name replacement with the CoNLL 2003 NER test data [27]. Sentences with more than 5 tokens (to ensure proper context) and contain exactly one unigram person entity (see limitations part of Section 4) are selected in this data. As a result, the sentence can have other n-gram entities of all types. This results in a dataset of 289 sentences. We again create synthetic sentences by replacing the unigram PERSON entity with the names described above.

Finally, we replicate our evaluations on lower-cased data (both Winogender and In-Situ) to investigate how the models perform when the sentences (including the names) are lower-cased; this removes the dominance of word shape features and checks purely for syntactic feature usage. This setting also resembles social media text, where capitalization rules are not very often followed [15, 16].

2.2 Models

We assessed the bias on the following widely-used NER model architectures as well as off-the-shelf libraries:

[4] While we are aware that name distributions might have changed slightly in recent years, we think it's a reasonable list for this project

[5] https://github.com/commonsense/conceptnet5/blob/master/conceptnet5/vectors/evaluation/bias.py

WINOGENDER	
Original	**$OCCUPATION** told **$PARTICIPANT** that **$NOM_PRONOUN** could pay with cash.
Sample 1	**Alya** told **Jasmine** that **Andrew** could pay with cash.
Sample 2	**Alya** told **Theo** that **Ryan** could pay with cash.
IN-SITU (CoNLL 03 Test)	
Original	**Charlton** managed Ireland for 93 matches , during which time they lost only 17 times in almost 10 years until he resigned in December 1995 .
Sample 1	**Syed** managed Ireland for 93 matches , during which time they lost only 17 times in almost 10 years until he resigned in December 1995 .

Table 2: Examples of synthetic dataset generated from Winogender Schema and CoNLL 03 test data.

(1) **BiLSTM CRF** [10, 12] is one of the most commonly-used deep learning architectures for NER. The model uses pre-trained word embeddings as input representations, bidirectional LSTM to compose context-dependent representations of the text from both directions, and Conditional Random Field (CRF) [11] to decode output into a sequence of tags. Since we are interested in both the correctness as well as the confidence of extracted named entities, we also compute the entity-level confidence via the *Constrained Forward-Backward algorithm* [5]. Different versions of this model were trained on CoNLL 03 NER benchmark dataset [27] by utilizing varying embedding methods:
   (a) **GloVe** uses GloVe 840B word vectors pre-trained on Common Crawl [19].
   (b) **CNET** uses ConceptNet english embeddings (version 1908) [25], which have already been debiased for gender and ethnicity [6].
   (c) **ELMo** uses ELMo embeddings [20], which provides contextualized representations from a deep bidirectional language model where the words are encoded using embeddings of their characters. This approach allows us to overcome the OOV issue.

(2) **spaCy** is a widely-used open-source library for NLP that features pre-trained NER models. We performed analysis on **spacy_sm** and **spacy_lg** English NER models from spaCy version 2.1.0[7]. spaCy models are trained on OntoNotes 5[8] data.

(3) **Stanford CoreNLP** [9] (**corenlp**) [13] is one of the most popular NLP library and we use the *2018-10-05* version. CoreNLP NER was trained (by its authors) on data from CoNLL03 and ACE 2002[10].

Note on excluding BERT While the approach of fine-tuning large pre-trained transformer language models such as BERT [6] has established state-of-the-art performance on NER, the implementations used subword tokenization such as WordPiece [28] or Byte-Pair-Coding [28] which require pre-tokenization followed by word pieces for NER tasks where the prediction has to be made on the word level. Although the BERT paper has addressed this issue by using the embedding of the first subword token for each word, this breaks the unigram entity assumption we have used in our analysis. Furthermore, number of BERT tokens may vary for names adding another degree of freedom to control. Furthermore, our inclusion of ELMo can be considered as a fair comparison for utilizing contextual word embeddings compared to other models which uses fixed word embeddings.

[6]https://blog.conceptnet.io/posts/2017/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/
[7]https://spacy.io/
[8]https://spacy.io/models/en
[9]https://stanfordnlp.github.io/CoreNLP/history.html
[10]https://nlp.stanford.edu/software/CRF-NER.html#Models

2.3 Evaluation Criteria

The goal of this work is to assess if NER models vary in their accuracy of identifying first names from various demographics as an instance of named entity with label $l = PERSON$. Assuming N_c unique names in a demographic category c, we define the metric $p_n^l = p(l|n)$ for each name n. We utilize this metric for our evaluations via various methods described below.

We first compare the overall accuracy of identifying names as person entity for each demographic category c. This is equal to $p_c^l = \sum_{n \in c} p(n) * p(l|n)$.

Next, we compare the distribution of accuracy across all the names of a given demographic. We compare the empirical cumulative density function (ECDF) of the accuracy p_n^l across all the names n for a given category c. This approach allows us to answer the question what percentage of names in a given category have an accuracy lower than x. We are particularly interested in observing what percentage of names in a category have an accuracy lower than the accuracy for the OOV name with uninformative embeddings.

In our final comparison, we utilize the confidence estimates of the model (whenever available) for entities which are predicted as person. For each name we compute the minimum, mean, median, and standard deviation of the confidence scores. We use these scores to identify the bias in the models.

3 RESULTS

3.1 Overall Accuracy

We describe the overall accuracy of various models across demographic categories in Table 3. We observe that the accuracy on White names (both male and female) is the highest (except for the ELMo model where the accuracy is highest for Muslim Male names) across all demographic categories and models. We also recognize

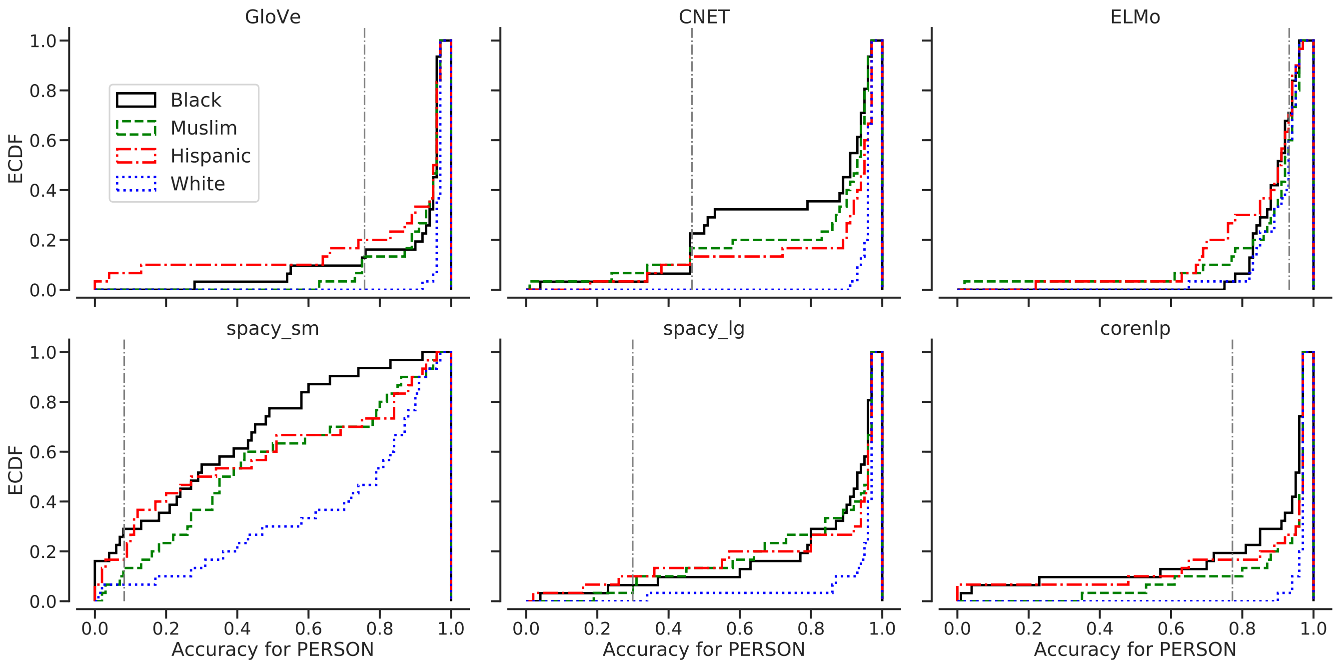


Figure 1: (Best viewed in color) Empirical Cumulative Density Function (ECDF) of names accuracy in Winogender data across demographic categories. The grey vertical line is the confidence percentile for OOV Name. Models with more left skewed accuracy are better (or harder to distinguish plots mean better models).

that the ELMo model exhibits the least variation in accuracy across all demographics, including the OOV names. For the ELMo model the bottom three names with the lowest accuracy are Jana (MF), Santiago (HM), and Salma (MF). Among these Jana and Santiago are also most likely to be identified as location entities while Salma is likely to be identified as person entity for 51% cases and location one for 36%.

We observe considerably lower accuracy (3%-30%) on uncapitalized names, particularly from the pre-trained CoreNLP and spaCy models such that the bias is no longer evident across the demographic groups (more details in table 7). Based on these low accuracy scores, we exclude the results of uncapitalized names in further sections. The above results indicate that all considered models are less accurate across non-White names. However, the character embedding based models like ELMo contain the least variation in accuracy across all demographics.

3.2 Distribution of Accuracy across Names

Next we look at the distribution of accuracy across names in each demographic category. In Figure 1, we report the distribution of name accuracy in Winogender data across all the names in a demographic category for all models. We observe that a large percentage of names from non-White categories have accuracy lower than the OOV names with uninformative embeddings. A similar analysis was conducted for all demographic categories (see figure 4) as well as only for gender categories (see figure 5), but the bias for gender is not as dominant as the other demographic categories. This indicates that the models introduce some biases based on the name's word vector, which causes the lower accuracy of these names. In table 4, we report the variation of accuracy across all names in a given demographic category and confirm that the ELMo model has the least variation. We observe similar results on the In-situ dataset (see figures 6, 8, and 7).

3.3 Model Confidence

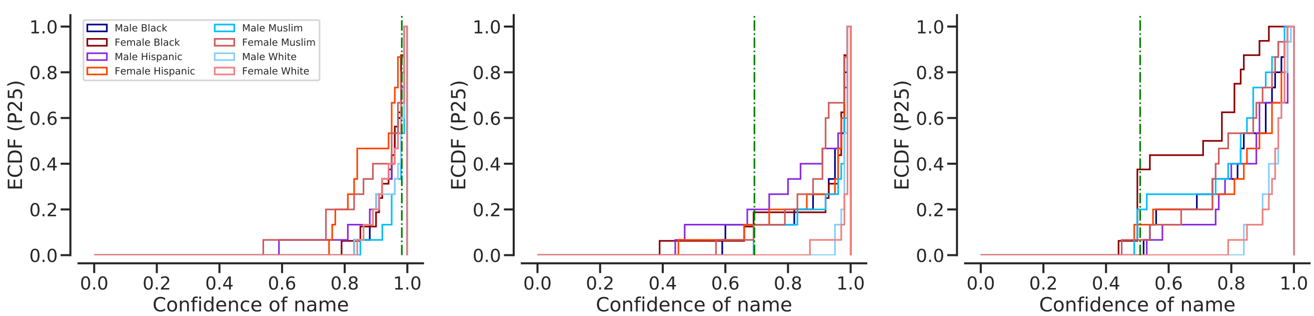
Finally, we investigate the distribution of model confidence across the names which were predicted as person. We use various percentile values for a given name's confidence. We analyze the 25th percentile confidence and the median confidence. As the percentile decreases, the bias observed should become more evident as it highlights the noisier tail of the data. In Figure 2, we report the distribution of the 25 percentile values. As before, we observe that a larger percentage of White names have a higher confidence compared to non-White names. Similarly, it can be observed that ELMo based models have the lowest variation in confidence values across all demographics. Surprisingly, the CNET models which are trained on debiased embeddings have the highest variation in confidence estimates. We investigate the variations in median confidence across names in each demographic in Table 5. This table confirms our observation above, that ELMo model has least variation across names. We again observe the similar trends for the in-situ data.

4 DISCUSSION

Our work sheds light on the variation in accuracy of named entity recognition systems on first names which are prominent, in certain demographic categories such as gender and race. A lower dimension

	CNET	ELMo	GloVe	corenlp	spacy_lg	spacy_sm
WINOGENDER						
Black Female	0.7039	0.8942	0.8931	0.7940	0.8908	0.3043
Black Male	0.8410	0.8986	0.9015	0.8862	0.7831	0.3517
Hispanic Female	0.8454	0.8308	0.8738	0.8626	0.8378	0.3726
Hispanic Male	0.8801	0.8603	0.7942	0.8629	0.8151	0.4628
Muslim Female	0.8537	0.8130	0.9074	0.8747	0.8287	0.4285
Muslim Male	0.7791	0.9265	0.9351	0.9477	0.8285	0.4976
White Female	0.9627	0.9116	0.9679	0.9723	0.9577	0.5574
White Male	0.9644	0.9068	0.9700	0.9688	0.9260	0.7732
OOV Name	0.4658	0.9318	0.7573	0.7724	0.2994	0.0824
IN-SITU						
Black Female	0.8289	0.8802	0.9193	0.8134	0.6732	0.2104
Black Male	0.8964	0.8800	0.9206	0.8828	0.5922	0.2651
Hispanic Female	0.8934	0.8510	0.9091	0.8754	0.6736	0.3038
Hispanic Male	0.9151	0.8729	0.8404	0.8699	0.6692	0.3649
Muslim Female	0.9015	0.8348	0.9230	0.8817	0.5686	0.3409
Muslim Male	0.8574	0.9043	0.9407	0.9421	0.6890	0.4122
White Female	0.9619	0.8900	0.9555	0.9714	0.7862	0.4503
White Male	0.9541	0.8930	0.9504	0.9589	0.7234	0.6388
OOV Name	0.7405	0.8962	0.8720	0.8374	0.1003	0.0381

Table 3: Overall accuracy for each demographic category, with highlighted best and worst performance. We observe significant performance gap between White names and names from other demographics.



(a) ELMo model **(b) GloVe model** **(c) CNET model**

Figure 2: (Best viewed in color) ECDF of percentiles of confidence values for a name to be identified as person entity. The vertical line is the confidence percentile for OOV name baseline.

projection (obtained via t-SNE) of the embeddings as shown in Figure 3) reveals that the name embeddings do cluster based on their demographic information. The clustering is more prominent across the race dimension.

It is important to note that the performance gap between names from different demographic groups can be partially attributed to the bias in the training data. Built from the Reuters 1996 news corpus, CoNLL03 is one of the most widely-used NER dataset. However, as shown in Table 6, the CoNLL03 training data contains significantly more Male names than Female names and more White names than non-White names.

While this work has approached studying the issue of bias using a synthetic dataset, it is still helpful in uncovering various aspects of the NER pipeline. We specifically identified variation in NER accuracy by using different embeddings. This is important because NER facilitates multiple automated systems, e.g. knowledge base construction, question answering systems, search result ranking,

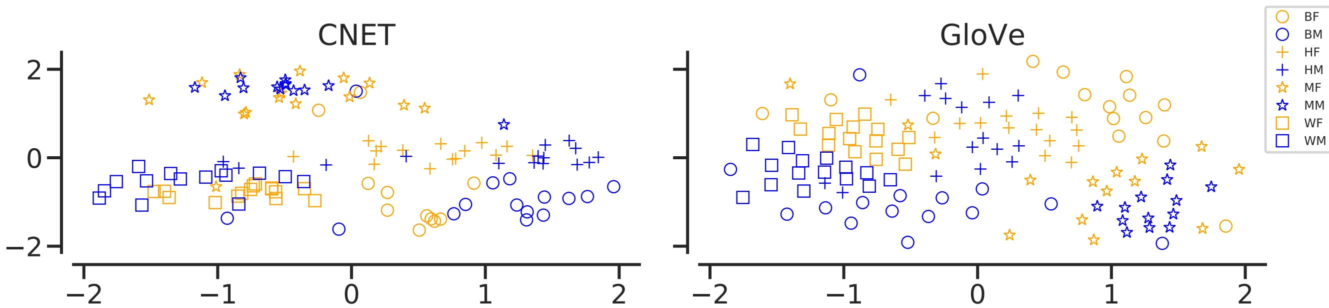


Figure 3: t-SNE projections of first name embeddings identified by their demographic categories (best viewed in color).

model	min⋆	mean⋆	std†	median⋆
		WINOGENDER		
CNET	0.02	0.846	0.223	0.948
GloVe	0.00	**0.903**	0.170	**0.965**
ELMo	**0.03**	0.881	**0.126**	0.922
corenlp	0.00	0.887	0.220	0.974
spacy_lg	0.00	0.847	0.241	0.965
spacy_sm	0.00	0.460	0.327	0.425
		IN-SITU		
CNET	0.242	0.898	0.130	0.952
GloVe	0.159	**0.919**	0.100	0.948
ELMo	**0.343**	0.876	**0.067**	0.889
corenlp	0.000	0.891	0.204	**0.969**
spacy_lg	0.000	0.662	0.255	0.775
spacy_sm	0.000	0.366	0.280	0.294

Table 4: Range of accuracy values across all names per demographic for each model. Lower is better for † and higher is better for ⋆.

model	min⋆	mean⋆	std†	median⋆
		WINOGENDER		
CNET	0.495	0.894	0.132	0.956
GloVe	0.468	0.952	0.104	0.994
ELMo	**0.621**	**0.980**	**0.046**	**0.995**
		IN-SITU		
CNET	0.606	0.946	0.080	0.981
GloVe	0.668	0.983	0.049	**0.998**
ELMo	**0.831**	**0.994**	**0.017**	**0.998**

Table 5: Range of median confidence values across all names per demographic for each model. Confidence values unavailable for other models. Lower is better for † and higher is better for ⋆

and automated keyword identification. If named entities from certain parts of the populations are systematically misidentified or mislabeled, the damage will be twofold: they will not be able to benefit from online exposure as much as they would have if they belonged to a different category (Allocation Bias [11] as defined in [1]) and they will be less likely to be included in future iterations of training data therefore perpetuating the vicious cycle (Representation bias). Furthermore, while a lot of research in bias has focused on just one aspect of demographics (i.e. only race or only gender) our work focuses on the intersectionality of both these factors. Similar research in the domain of bias across gender, ethnicity, and nationality has been studied in bibliometric literature [17].

Limitations Our current work is limited in its analysis to only unigram entities. A major challenge for correctly constructing and evaluating our methods for n-gram entities is to come up with a collection of names which are representative of demographics. While first name data is easily available through various census portals, full name data tagged with demographic information is harder to find. Furthermore, when extending this analysis to n-gram entities we need to define better evaluation metrics, i.e. how different is a mistake on the first name from a mistake on other parts of the name, and how to quantify this bias appropriately. Finally, we are aware that our name lists are based on old data and certain first names are be more likely to be adopted by other communities, leading to the demographic association of names to change across time [24]. However, these factors do not affect our analysis as our name collection consists of dominant names in a demographic. Additionally, our work can be extended to other named entity categories like location, and organizations from different countries so as to assess the bias in identifying these entities. Since, our analysis focused on NER models trained on English corpus, another line of research will be to see if models trained in other languages also contain favorable results for named entities more likely to be used in cultures where that language is popular. This should lead to the assessment of NER models in different languages with named entities representing a larger demographic diversity. Finally, the goal of this paper has been to identify biases in accuracy of NER models. We are investigating ways to mitigate these biases in an efficient manner.

[11] https://catalogofbias.org/biases/allocation-bias/

5 RELATED WORK

Bias in embeddings has been studied by Bolukbasi et al. [3], who showed that the vector for stereotypically male professions are closer to the vector for "man" than "woman" (e.g. "Man is to Computer Programmer as Woman is to Homemaker"). Techniques to debias embeddings were suggested, where a "gender" direction is identified in the vector space and thus subtracted from the embeddings. More recently Gonen and Goldberg [7] showed how those efforts are not substantially removing bias, rather hiding it: words with similar biases are still clustered together in the de-biased space. Manzini et al. [14] extended the techinques of [3] to multi-class setting, instead of just binary ones. Emebddings were also the subject of scrutiny in Caliskan et al. [4], where a modified version of the implicit association tests [8] were developed.

The Winogender schemas we used in this works were developed by [22] to study gender bias in coreference resolution.

6 CONCLUSION

In this work, we introduced a novel framework to study the bias in named entity recognition models using synthetically generated data. From our analysis reports that models are better at identifying White names across all datasets with higher confidence compared with other demographics such as Black names. We also demonstrate that debiased embeddings do not help in resolving the bias in recognizing names. Finally, our results show that character based models, such as ELMo, result in the least bias across demographic categories, but those models are still unable to entirely remove the bias. Since, NER models are often the first step in automatic construction of knowledge bases, our results can help identify potential issues of bias in KB constructions.

ACKNOWLEDGMENTS

REFERENCES

[1] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In *Proceedings of the 9th Annual Conference of the Special Interest Group for Computing, Information and Society* (Philadelphia, PA, USA). http://meetings.sigcis.org/uploads/6/3/6/8/6368912/program.pdf

[2] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (September 2004), 991–1013. https://doi.org/10.1257/0002828042002561

[3] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer As Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., USA, 4356–4364. http://dl.acm.org/citation.cfm?id=3157382.3157584

[4] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356 (2017), 183–186.

[5] Aron Culotta and Andrew McCallum. 2004. Confidence Estimation for Information Extraction. In *Proceedings of HLT-NAACL 2004: Short Papers*. Association for Computational Linguistics, Boston, Massachusetts, USA, 109–112. https://www.aclweb.org/anthology/N04-4028

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 609–614. https://doi.org/10.18653/v1/N19-1061

[8] Anthony G Greenwald, Debbie E. McGhee, and Joe L. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74 6 (1998), 1464–80.

[9] Weiwei Guo, Huiji Gao, Jun Shi, and Bo Long. 2019. Deep Natural Language Processing for Search Systems. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. ACM, New York, NY, USA, 1405–1406. https://doi.org/10.1145/3331184.3331381

[10] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *ArXiv* abs/1508.01991 (2015).

[11] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289. http://dl.acm.org/citation.cfm?id=645530.655813

[12] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 260–270. https://doi.org/10.18653/v1/N16-1030

[13] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, 55–60. https://doi.org/10.3115/v1/P14-5010

[14] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *NAACL-HLT*.

[15] Shubhanshu Mishra. 2019. Multi-Dataset-Multi-Task Neural Sequence Tagging for Information Extraction from Tweets. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media* (Hof, Germany) *(HT '19)*. Association for Computing Machinery, New York, NY, USA, 283–284. https://doi.org/10.1145/3342220.3344929

[16] Shubhanshu Mishra and Jana Diesner. 2016. Semi-supervised Named Entity Recognition in noisy-text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. The COLING 2016 Organizing Committee, Osaka, Japan, 203–212. https://www.aclweb.org/anthology/W16-3927

[17] Shubhanshu Mishra, Brent D. Fegley, Jana Diesner, and Vetle I. Torvik. 2018. Self-citation is the hallmark of productive authors, of any gender. *PLOS ONE* 13, 9 (sep 2018), e0195773. https://doi.org/10.1371/journal.pone.0195773

[18] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism* (first ed.). New York University Press.

[19] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[20] Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1756–1765. https://doi.org/10.18653/v1/P17-1161

[21] Jay Pujara and Sameer Singh. 2018. Mining Knowledge Graphs From Text. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. ACM, New York, NY, USA, 789–790. https://doi.org/10.1145/3159652.3162011

[22] Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social Bias in Elicited Natural Language Inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 74–79. https://doi.org/10.18653/v1/W17-1609

[23] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 8–14. https://doi.org/10.18653/v1/N18-2002

[24] Brittany N. Smith, Mamta Singh, and Vetle I. Torvik. 2013. A Search Engine Approach to Estimating Temporal Changes in Gender Orientation of First Names. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (Indianapolis, Indiana, USA) *(JCDL '13)*. ACM, New York, NY, USA, 199–208. https://doi.org/10.1145/2467696.2467720

[25] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, California, USA) *(AAAI'17)*.

AAAI Press, 4444–4451. http://dl.acm.org/citation.cfm?id=3298023.3298212

[26] Chris Sweeney and Maryam Najafian. 2019. A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1662–1667. https://doi.org/10.18653/v1/P19-1162

[27] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 142–147. https://www.aclweb.org/anthology/W03-0419

[28] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR* abs/1609.08144 (2016). arXiv:1609.08144 http://arxiv.org/abs/1609.08144

A APPENDIX

A.1 Name distribution in data

Category	Total Count	Most Common Name (Count)
Black Female (BF)	0	–
Black Male (BM)	18	Malik (13)
Hispanic Female (HF)	22	Maria (12)
Hispanic Male (HM)	89	Jose (20)
Muslim Female (MF)	8	Jana (6)
Muslim Male (MM)	68	Ahmed (49)
White Female (WF)	17	Stephanie (6)
White Male (WM)	148	Paul (51)

Table 6: Name distribution in CoNLL03 training data across different categories

A.2 Distribution of accuracy for various subsets of data for Winogender analysis

	CNET	ELMo	GloVe	corenlp	spacy_lg	spacy_sm
WINOGENDER LOWER						
Black Female	0.0018	0.8695	0.6855	0.0230	0.0915	NaN
Black Male	0.0911	0.8764	0.8068	0.0292	0.2077	NaN
Hispanic Female	0.0572	0.8137	0.7624	0.0581	0.1496	NaN
Hispanic Male	0.0556	0.8401	0.7408	0.0321	0.3044	NaN
Muslim Female	0.0192	0.7982	0.7517	0.0164	0.1797	NaN
Muslim Male	0.0222	0.9031	0.8118	0.0088	0.2787	NaN
White Female	0.0288	0.8779	0.8363	0.0552	0.1385	0.0000
White Male	0.0318	0.8736	0.7839	0.0193	0.2920	NaN
OOV Name	NaN	0.9256	0.0001	NaN	NaN	NaN
IN-SITU LOWER						
Black Female	0.0087	0.8774	0.7855	0.0151	0.0519	NaN
Black Male	0.1679	0.8759	0.8895	0.0291	0.0877	NaN
Hispanic Female	0.1066	0.8482	0.8750	0.0678	0.0634	NaN
Hispanic Male	0.1137	0.8697	0.8226	0.0429	0.1712	NaN
Muslim Female	0.0480	0.8332	0.8706	0.0136	0.1045	NaN
Muslim Male	0.0544	0.8987	0.8517	0.0065	0.1453	NaN
White Female	0.0826	0.8844	0.9340	0.0544	0.0388	0.0005
White Male	0.0867	0.8872	0.9059	0.0418	0.1398	NaN
OOV Name	NaN	0.8962	0.2353	NaN	NaN	NaN

Table 7: Overall accuracy on lower cased data for each demographic category, with highlighted best and worst performance.

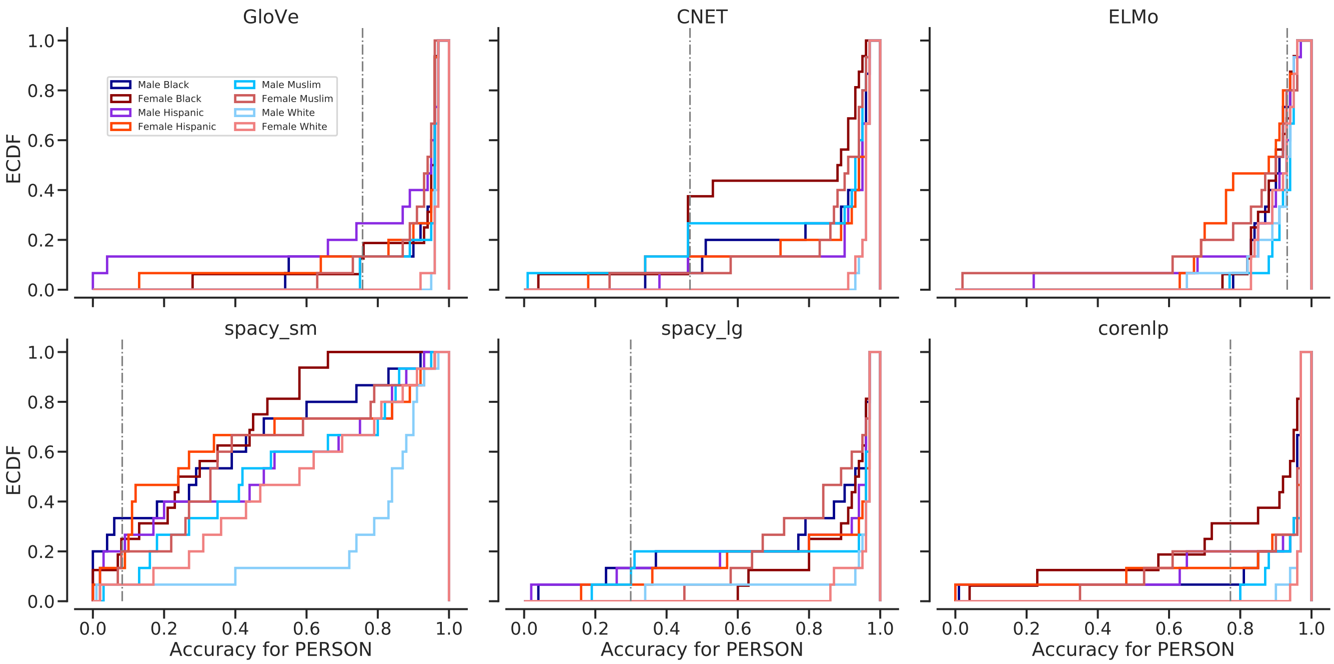


Figure 4: (Best viewed in color) Empirical Cumulative Density Function (ECDF) of names accuracy in Winogender data across demographic categories. The grey vertical line is the confidence percentile for OOV Name. Models with more left skewed accuracy are better (or harder to distinguish plots mean better models).

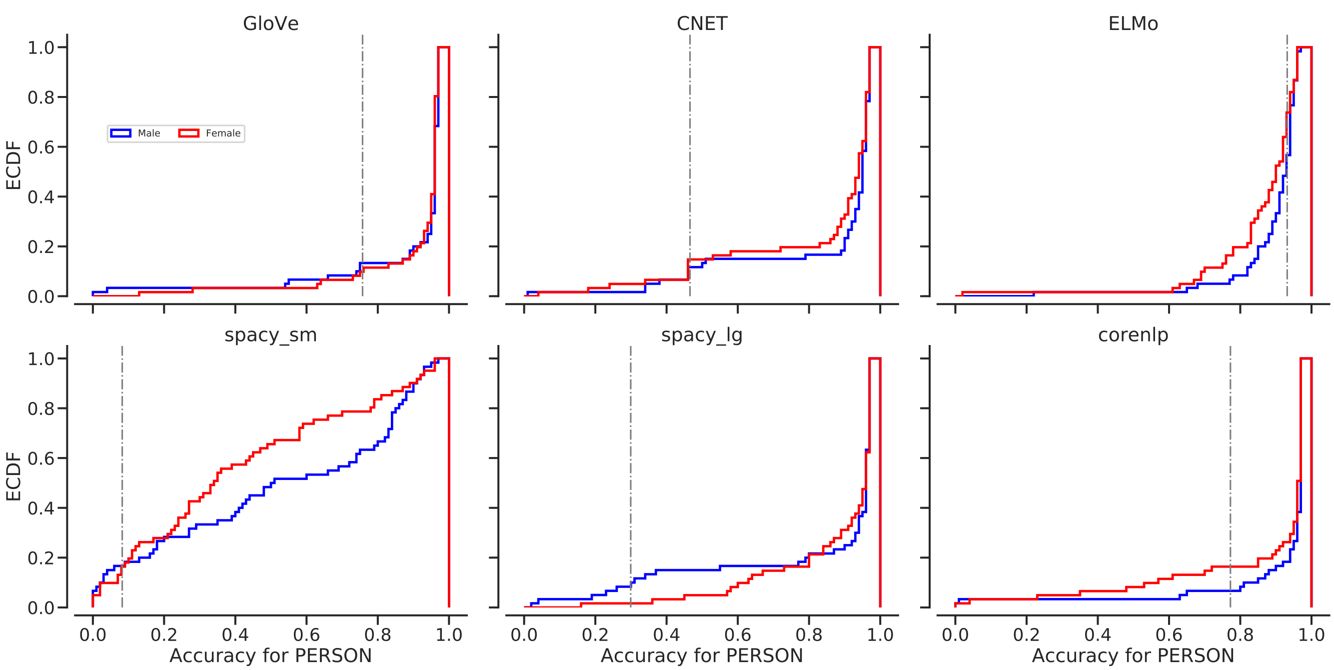


Figure 5: (Best viewed in color) Empirical Cumulative Density Function (ECDF) of names accuracy in Winogender data across demographic categories. The grey vertical line is the confidence percentile for OOV Name. Models with more left skewed accuracy are better (or harder to distinguish plots mean better models).

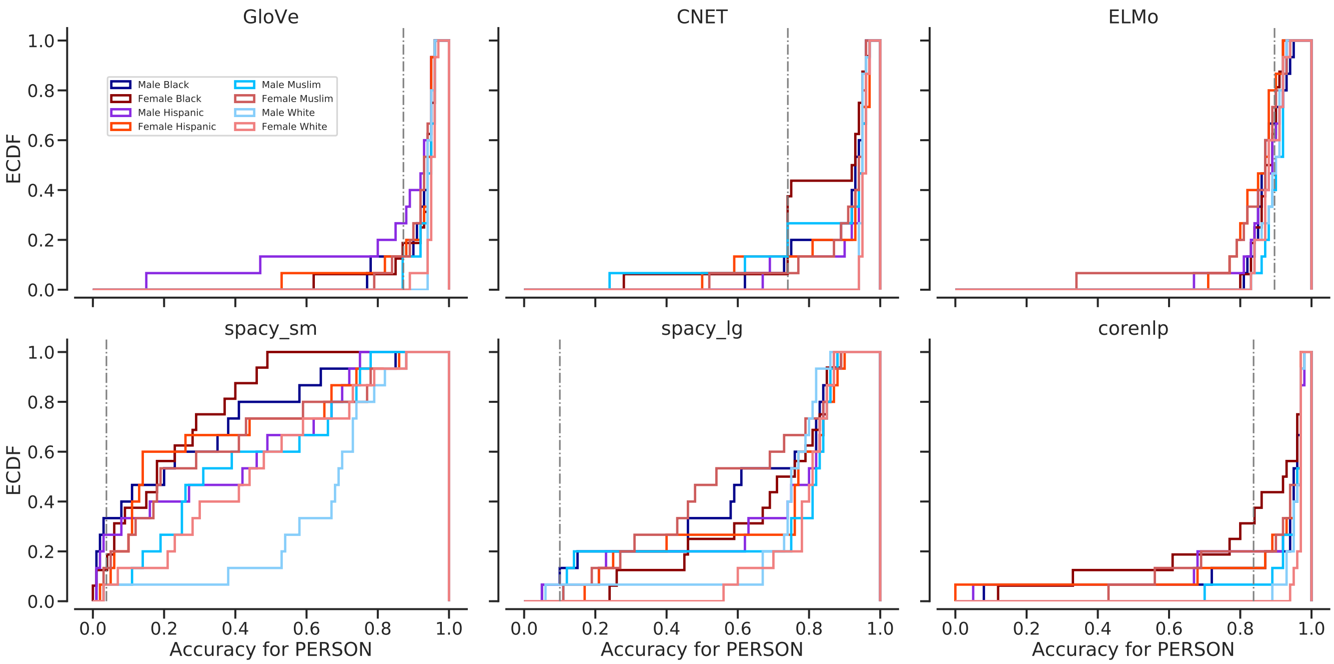


Figure 6: (Best viewed in color) Empirical Cumulative Density Function (ECDF) of names accuracy in In-Situ data across demographic categories. The grey vertical line is the confidence percentile for OOV Name. Models with more left skewed accuracy are better (or harder to distinguish plots mean better models).

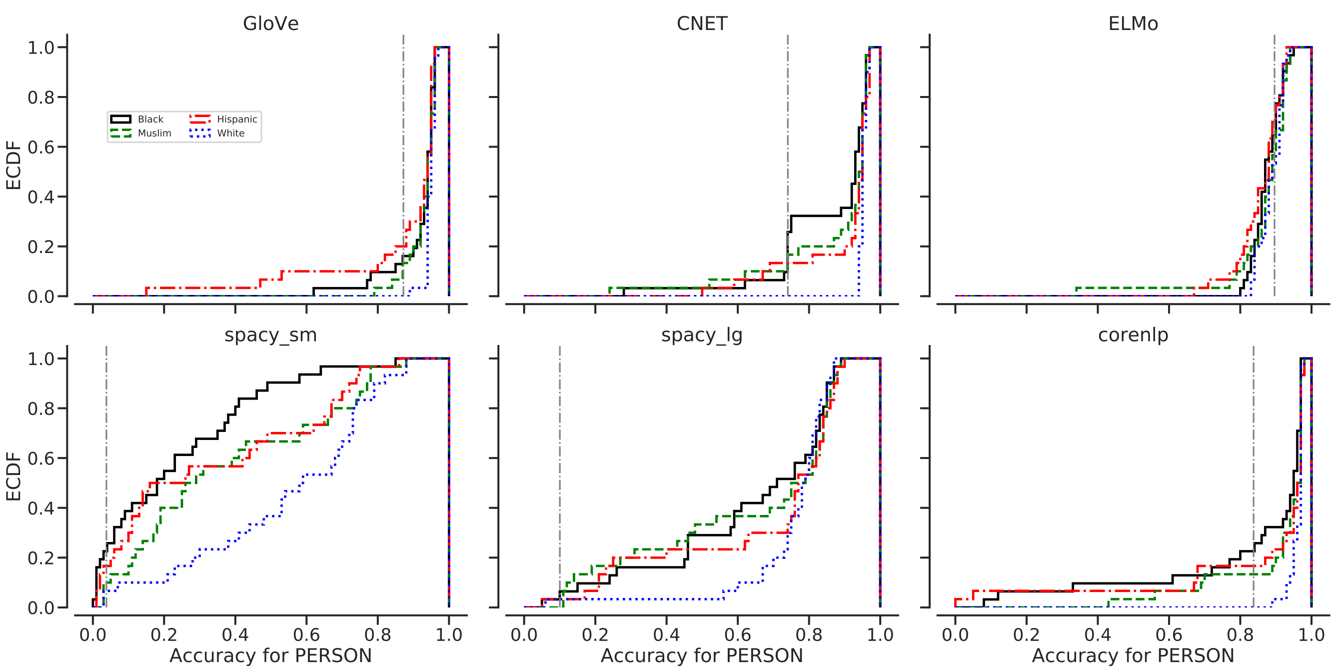


Figure 7: (Best viewed in color) Empirical Cumulative Density Function (ECDF) of names accuracy in In-Situ data across demographic categories. The grey vertical line is the confidence percentile for OOV Name. Models with more left skewed accuracy are better (or harder to distinguish plots mean better models).

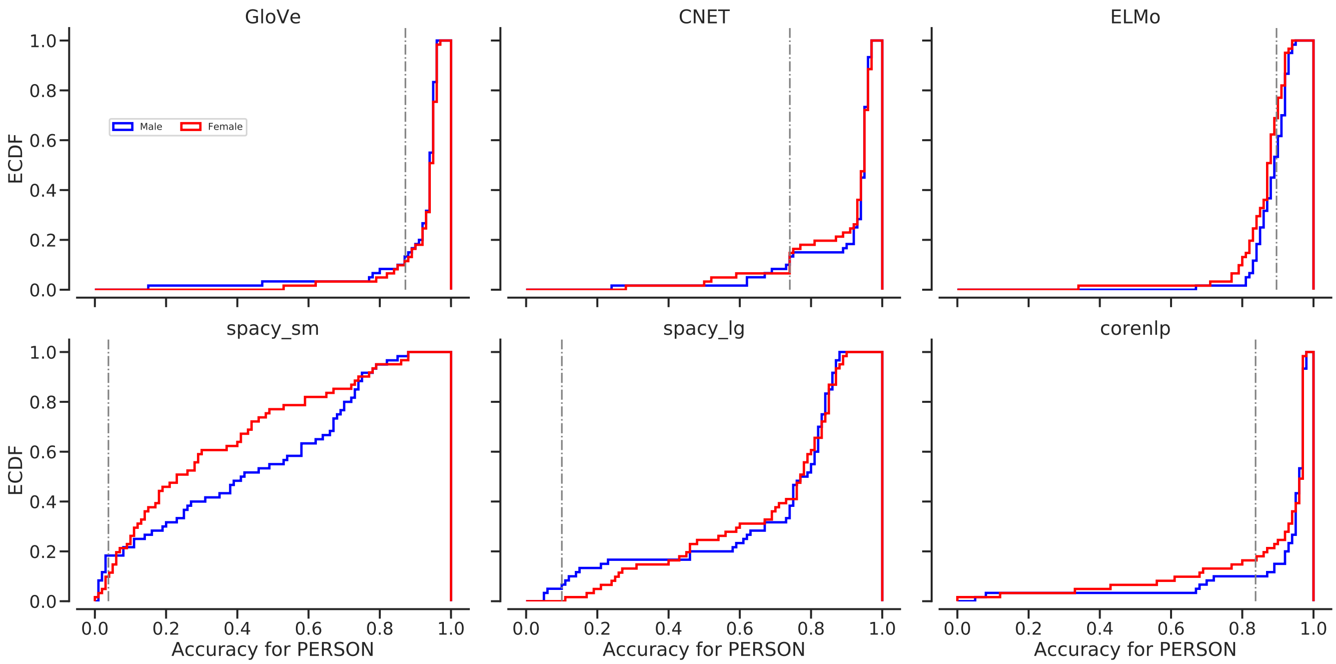


Figure 8: (Best viewed in color) Empirical Cumulative Density Function (ECDF) of names accuracy in In-Situ data across demographic categories. The grey vertical line is the confidence percentile for OOV Name. Models with more left skewed accuracy are better (or harder to distinguish plots mean better models).