Behind the Mask: Demographic bias in name detection for PII masking

Courtney Mansfield*, **Amandalynne Paullada***†, **Kristen Howell***
*LivePerson Inc., Seattle, Washington, USA
†Biomedical Informatics & Medical Education, University of Washington, Seattle, Washington, USA
cmansfield@liveperson.com, paullada@uw.edu, khowell@liveperson.com

Abstract

Many datasets contain personally identifiable information, or PII, which poses privacy risks to individuals. PII masking is commonly used to redact personal information such as names, addresses, and phone numbers from text data. Most modern PII masking pipelines involve machine learning algorithms. However, these systems may vary in performance, such that individuals from particular demographic groups bear a higher risk for having their personal information exposed. In this paper, we evaluate the performance of three off-the-shelf PII masking systems on name detection and redaction. We generate data using names and templates from the customer service domain. We find that an open-source RoBERTa-based system shows fewer disparities than the commercial models we test. However, all systems demonstrate significant differences in error rate based on demographics. In particular, the highest error rates occurred for names associated with Black and Asian/Pacific Islander individuals.

1 Introduction

In a time of extensive data collection and distribution, privacy is a vitally important but elusive goal. In 2021, the US-based Identity Theft Resource Center reported a 68% increase in data breaches from the previous year, with 83% involving sensitive information[1]. The exposure of personally identifiable information (PII), such as names, addresses, or social security numbers, leaves individuals vulnerable to identity theft and fraud. In response, a growing number of companies provide data protection services, including PII detection, redaction (masking), and anonymization.

PII masking offers assurances of security. However, this paper considers whether the models powering these services perform fairly across individuals, regardless of race, ethnicity, and gender. Historically, the US "Right to Privacy" concept has been centered around Whiteness, initially to protect White women from the then-emergent technology of photography and visual media (Osucha, 2009). Black individuals have had less access to privacy and face greater risk of harm due to surveillance, including algorithmic surveillance (Browne, 2015; Fagan et al., 2016).

In this paper, we evaluate the detection and masking of names, which are the primary indexer of a person's identity. We sample datasets of names and demographic information to measure the performance of off-the-shelf PII maskers. Although model bias or unfairness can be the result of a number of factors, including training data or presuppositions encoded in the algorithms themselves, the commercial systems we examine fail to provide details about training data or implementation. Therefore, we do not hypothesize a causal relationship between these factors and our findings.

Our work quantifies disparities in the name detection of PII masking systems where poor performance can directly and negatively impact individuals. We demonstrate significant disparities in the recognition of names based on demographic characteristics, especially for names associated with Black and Asian/Pacific Islander groups.

2 PII Masking

This study analyzes personally identifiable information (PII) masking systems which aim to detect and redact sensitive personal information, particularly names, from text. This has been an important problem in the biomedical domain, in terms of preparing de-identified patient data for research (Kayaalp, 2018), but is also increasingly important in an age of language models trained from web-

[1] https://www.idtheftcenter.org/post/identity-theft-resource-center-2021-annual-data-breach-report-sets-new-record-for-number-of-compromises/

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 76 - 89

scraped data, which have been shown to reveal private information that was not removed from the underlying training data (Carlini et al., 2021).

Since early efforts masking data by hand, automated methods have been employed, from using word lists or dictionaries (Thomas et al., 2002), which do not generalize to unseen names and locations, to rule-based or regular expression systems (Beckwith et al., 2006; Friedlin and McDonald, 2008), which are generalizable, but can be brittle. These have been replaced with machine learning systems (Szarvas et al., 2006; Uzuner et al., 2008) and most recently neural networks (Dernoncourt et al., 2017; Adams et al., 2019).

Modern PII maskers rely on Named Entity Recognition (NER) to identify entities (e.g. name and location) for redaction. NER has had recent success with hybrid bi-directional long short term memory (BiLSTM) and conditional random field (CRF) models (Huang et al., 2015), and following the general trend in NLP, fine-tuning on large language models such as BERT (Li et al., 2019). Additional discussion on NER architectures can be found in Li et al. (2020).

Previous research in Named Entity Recognition (NER) has illuminated race and gender-based disparities. Mishra et al. (2020) evaluates a number of NER models which consider performance according to gender and race/ethnicity. The analysis considers 15 names per intersectional group, finding that White-associated names are more likely to be recognized across all systems. Our work differs from and extends this work in key aspects: focusing on off-the-shelf PII masking, providing analysis on over 4K names, and reporting on significance and additional metrics.

Recent PII masking models perform extremely well in certain contexts. The recurrent neural network of Dernoncourt et al. (2017) achieves 99% recall overall and just below 98% for names on patient discharge summaries in the medical domain. The commercial models we consider do not advertise performance metrics, and as shown in Section 7, do not achieve such high performance across our datasets.

It is important to note that removing names alone is insufficient to fully protect individuals from being identified from data. Data sets can still reveal just enough information to re-identify individuals, as in the case of Massachusetts Governor William Weld, whose medical records, although not connected directly to his name in a de-identified data set, were traceable back to him by matching information from an easily attained external data resource (Sweeney, 2002). Here we focus on names as they are a primary identifier for an individual.

3 What's in a Name?

The primary goal of this paper is to understand whether, and to what degree, the performance of PII masking models is influenced by correlates of race, ethnicity, and gender. We frame bias in terms of significant discrepancies in performance based on race/ethnicity and gender, looking specifically to instances where private information was not masked (false negative rates, described in Section 6.2). PII masking is a primary mechanism for protecting personal data, and a systematic failure to mask information belonging to marginalized subgroups can cause undue harm to those populations, through identity theft, identity fraud, and loss of privacy. Names are not a proxy for gender or race/ethnicity, but our rationale is as follows: if most of the people with Name N have self-identified as belonging to Group G_1, and Name N is frequently miscategorized by PII systems at a rate that is higher than that for a name more commonly used by individuals in Group G_2, then we argue that members of Group G_1 bear a higher privacy risk.

We focus our analysis on given names (sometimes known as 'first names') and family names (sometimes known as 'surnames' or 'last names'). Naming conventions vary in different cultural and linguistic contexts. In many cultures, given names and/or family names can be gendered, or disproportionately associated with a particular gender, religious or ethnic group. In the present study, gender, race and ethnicity are considered with respect to a defined set of categories for the purpose of analysis, but we acknowledge that such labels are socially constructed and mutable over time and space (Sen and Wasow, 2016).

Previous research has uncovered racial and gender discrimination based on individual names. Bertrand and Mullainathan (2004) found that, given identical resumes with only a change in name, resumes with Black-associated names received fewer callbacks than White-associated names. Sweeney (2013) found that internet searches for Black (in contrast to White) names were more likely to trigger advertisements that suggested the existence of arrest records for people with those names.

We do not attempt to infer personal information tied to names in our data, but rather, rely on real, self-reported information. However, there are limitations to using standardized gender and racial categories in studying algorithmic fairness, even when individuals are able to self-identify (Hanna et al., 2020). Within each racial/ethnicity category made available on the standardized forms in the data we use (described in Section 4), for example, there is a large variety in the linguistic cultures and naming practices encompassed in each group. Our intent is not to conflate race and ethnicity and language, but rather to get a coarse-grained look at performance of PII masking systems on names that are strongly associated with the demographic groupings that are available. Similarly, the available data limits gender categories to the binary 'male' and 'female,' and while names are not a good proxy for gender, we look for strong associations in the data, as described further in Section 4.

4 Data

In this section, we describe our method for creating test sentences for evaluating name detection in PII masking models. In our evaluation, we use a sentence perturbation technique which is employed in previous studies to test model performance across sensitive groups (Garg et al., 2019; Hutchinson et al., 2020). Using a variety of templates, we fill slots with names from the datasets, allowing us to measure performance across race/ethnicity and gender.

Reliable sources of demographically labeled names are difficult to find and using real names is an issue of privacy. Therefore, we consider datasets of names with aggregate demographic information as a proxy. We also evaluate on the names of US Congress members, whose identity and self-reported demographic information is publicly available. Templates and source datasets are described in the following sections.

4.1 Templates

We collected a set of 32 templates from real-world customer service messaging conversations (see examples in Table 1 and the full set in Appendix A.3). These include dialog between customers and conversational AI or human agents. Customer service data is especially vulnerable to security threat, carrying potentially sensitive personal information such as credit card or social security numbers. Topics of discussion in the dataset include placing or tracking a purchase or paying a bill. Each template contains a name, which we replace with a generic NAME slot. Various identifiers from the dataset (e.g. location or reference numbers) are swapped to protect personal information.

Sample Templates
This was from <NAME>
The response is signed <NAME>
it's YGDFEA the reservation. <NAME>

Table 1: Sample of templates used for analysis.

4.2 LAR Data

The LAR dataset from Tzioumis (2018) contains aggregate names with self-reported race/ethnicity from US Loan Application Registrars (LARs). It includes 4.2K given names from 2.6M observations across the US. Race/ethnicity categories are shown in Table 2.

There are limitations to the Tzioumis (2018) dataset. Because the sample is drawn from mortgage applications and there are known racial and socioeconomic differences in who applies for mortgage applications (Charles and Hurst, 2002), the data is likely to contain representation bias. However, the LAR dataset is the largest available set of names and demographics, estimated to reflect 85.6% of names in the US population (Tzioumis, 2018). Due to its large size, we are able to control for the frequency of names, as described in Section 5.

4.3 NYC Data

The NYC dataset was created using the New York City (NYC) Department of Health and Mental Hygiene's civil birth registration data (NYC Open Data, 2013) and contains 1.8K given names from 1.2M observations. Data is available from 2011-2018 and includes self-reported race/ethnicity of the birth mother (other parents' information is not available). The sex of the baby is included, which permits an intersectional analysis.[2] The race/ethnicity groups are shown in Table 2.

While the other datasets report on adult names, the NYC data aggregates the names of children

[2] Although the NYC data includes the child's *sex assigned at birth*, we use this variable to approximate the *gender* associated with the name.

who are between 4-11 at the time of this writing. This adds diversity in terms of age, as data privacy is an important issue for both children and adults.

4.4 Congress Data

The Congress dataset allows for evaluation over the given and family names of real individuals. The 540 current members of US Congress provide self-reported demographic information.[3] Race/ethnic groups are described in Table 2. 76% of congress members do not report membership in the race/ethnicity groups listed, and are grouped as "White/Other".

This dataset provides a naturalistic analysis of full names. Alternatively, one could programmatically generate given and family name pairs from datasets of first names and a dataset of last names. However, the broad race/ethnic groups used for classification do not account for the variance in the cultural backgrounds of the names (e.g. Pakistani and Native Hawaiian backgrounds are listed under the umbrella of Asian and Pacific Islander).

5 Sampling Process

This section describes the process of sampling the source names. The LAR and NYC datasets aggregate name counts and frequencies per race/ethnicity. We sample names which have a strong 'association' with a particular race/ethnicity and gender. Because frequency (i.e. popularity) of a name could contribute to spurious performance disparities between groups, we sample the LAR data so that all names are frequency matched across groups.

5.1 Demographic categorization

For each group, we sample names that are "associated" with that particular group. We define "association" as when 75% of people with the same name self-report within the same race/ethnicity. In the LAR dataset, the NH American Indian or Alaska Native and NH Multi-race names reflect 1% of individuals in the dataset (Tzioumis, 2018). No names were found with strong associations in these groups, and for this reason, we do not include them in the analysis. We map race/ethnicity groups across datasets to a common set of labels, which are based on categories of the 2010 US Census dataset of surname and race/ethnicity information (Comenetz, 2016). Race/ethnicity categorization for all datasets is shown in Table 2.

The NYC dataset also includes gender. Using a 90% threshold for our definition of 'association', 99% of names in the source set are strongly association with one gender.

5.2 Frequency matching

Because the LAR dataset has a large sample size, it is possible to control for the frequency of names while maintaining a minimum threshold of 20 names per category. To standardize based on frequency, we use counts from the 2010 US Census Bureau. We did not use observation counts directly from the LAR data, due to the aforementioned potential for representational bias.

We sample the LAR dataset to align the mean observation counts of Black-associated names and other groups, as there are few Black-associated names in the dataset (n=21). However, there is limited overlap in the frequency distributions of API-associated names with Hispanic and Black-associated names. Therefore, we sample a second set with API and White-associated names only. We refer to these datasets as LAR1 (Black, Hispanic, and White) and LAR2 (API and White). The frequency matching process is described in more detail in Appendix A.2.

6 Experiment Setup

The following sections discuss the PII masking systems we evaluate. We use several metrics to investigate the PII masking performance across name subsets.[4]

6.1 Models

We select two commercial and one open-source PII masking system for evaluation. The commercial systems we consider are Amazon Web Services (AWS) Comprehend and Google Cloud Platform Data Loss Prevention (GCP DLP). We choose these systems for their potentially large reach, with AWS and GCP holding a combined 43% market share of cloud services.[5] Amazon Comprehend provides an English model with a NAME entity for PII redaction. GCP DLP offers redaction and includes a

[3] See www.senate.gov and https://pressgallery.house.gov/member-data/demographics.

[4] Experiment code is publically available at https://github.com/csmansfield/pii-masking-bias.

[5] https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/

Data	Dataset Race/Ethnicity Group	Mapped label
LAR	NH Asian or Native Hawaiian or Other Pacific Islander	Asian and Pacific Islander
	NH Black or African American	Black
	Hispanic or Latino	Hispanic
	NH American Indian or Alaska Native	Indigenous
	NH Multi-race	Multi-race
	NH White	White
NYC	Asian and Pacific Islander	Asian and Pacific Islander
	Black	Black
	Hispanic White	Hispanic
	NH White	White
Cong.	Asian	Asian and Pacific Islander
	Black	Black
	Hispanic	Hispanic
	Indigenous	Indigenous
	White/Other	White

Table 2: Race/ethnicity categories used for each data source and the mapped set of race/ethnic group labels each category is mapped to for our analysis. The term "Non-Hispanic" is abbreviated NH.

global PERSON_NAME entity. Microsoft's Presidio is an open-source service for PII detection. We use the default English model which uses logic such as regex matching and Named Entity Recognition (NER). For the Presidio model we use a spaCy 3.2 en_core_web_trf model for NER, which utilizes the RoBERTa-base Transformer model trained on OntoNotes 5.

6.2 Evaluation metrics

We measure false negative rates (FNRs), the rate at which a PII system does not detect a name that is present in the dataset (and therefore is unable to mask it).[6] Following Dixon et al. (2018) we report on the False Negative Equality Difference, which measures differences between the false negative rate over the entire dataset and across each demographic subgroup g. We add a normalization term to compare the FNED of datasets with different numbers of groups, as shown in equation 1.

$$\frac{1}{|G|}\sum_{g\in G}|FNR - FNR_g| \tag{1}$$

We also measure the statistical significance of performance differences across subgroups. We conduct Friedman and Wilcoxon signed-rank tests following Czarnowska et al. (2021). The Friedman test is used for cases with more than 2 subgroups, and provides a single p-value for each dataset and system pair. The p-value determines whether to reject the null hypothesis that FNR of a given system is the same across all demographic groups. The statistic is calculated considering j demographic subsets g. First, we calculate the average FNR for a template t, over all names belonging to a particular subset g. The averages for each of the 32 templates considering group g are contained in X_g. The Friedman statistic is calculated for all X_g.

$$X_g = (FNR(x_g^1), ..., FNR(x_g^{32}))$$
$$Friedman(X_1, ..., X_j) \tag{2}$$

Nemenyi post-hoc testing is used for further pairwise analysis. For cases with only 2 subgroups, we alternatively perform Wilcoxon signed-rank tests. In order to control for multiple comparisons, we apply a Bonferroni correction across all p-values (at p<0.05 and n=15, our adjusted significance threshold is 0.003).

7 Results

We present the results of the evaluation, considering overall performance and performance related to race/ethnicity, gender, and intersectional factors. The section concludes with an analysis of errors.

[6] Whereas false positive rates are useful for evaluating the precision of a model, our focus is the failure to detect person names, rather than the incorrect identification of tokens that are not person names. Furthermore, we report no false positives in our findings.

	Group	N	FNR (%)		
			AWS	GCP	MP
LAR1	Black	20	20.0	18.1	**29.5**
	Hisp.	172	**28.4**	12.4	24.7
	White	1000	21.3	**18.5**	20.0
	All	1192	22.3	17.6	20.8
LAR2	API	441	**38.2**	**51.2**	**29.2**
	White	1000	25.3	18.6	25.8
	All	1441	29.3	28.6	26.8
NYC	API	165	21.3	43.6	22.0
	Black	226	**28.9**	**56.3**	**32.6**
	Hisp.	389	20.1	34.2	21.2
	White	592	26.9	29.2	25.9
	All	1359	24.6	36.8	25.2
Cong.	API	16	**23.0**	**12.1**	**11.7**
	Black	56	15.2	9.7	9.5
	Hisp.	48	13.9	8.3	9.4
	Indig.†	3	7.0	6.3	7.8
	Multi.†	6	8.3	6.3	10.9
	White/ Other	419	12.1	6.7	7.7
	All	530	12.8	7.3	8.1

Table 3: Support and average false negative rate (FNR) by race/ethnicity group across datasets. Groups marked with '†' are not included in formal statistical analysis due to low support. Maximum FNR per dataset/sytem is shown in bold.

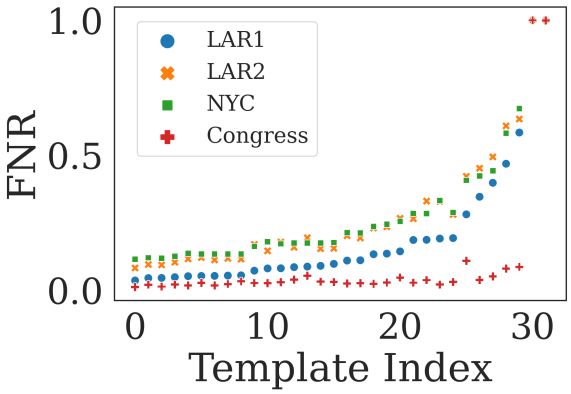


Figure 1: Average FNR across each template per dataset.

		FNED		
		AWS	GCP	MP
Race/ ethnicity	LAR1	*3.1	*2.2	***4.4**
	LAR2	***6.4**	***16.3**	1.7
	NYC	*3.6	*8.9	*3.7
	Congress	*3.6	*2.2	1.7
Gender	NYC	*3.2	*4.4	0.8
	Congress	*1.3	*0.6	0.2

Table 4: The normalized false negative equality difference (FNED) for race/ethnicity and gender subsets of the data. Asterisks indicate significance ($p<0.003$) in FNR differences by group. Maximum FNED per system is shown in bold.

7.1 Overall Performance

The average performance on the datasets can be seen in Table 3. System performance varies according to the dataset, with no single system performing best on all sets. All systems have lower FNR on the Congress dataset, where both given and family names are available, likely due to the increased information load of full names. The LAR2 and NYC names prove the most challenging across all systems.

The average performance of the names per each template is shown in Figure 1. Performance varies considerably, with average FNR per template ranging between 6%. and 100%. The mean FNR for all templates is 22%.

7.2 Performance by Race/Ethnicity

The normalized false negative equality differences (FNEDs) are shown in Table 4.

The highest FNED, which is an 82% increase over the second highest FNED, is seen in GCP's performance over the LAR2 dataset which includes frequency controlled API and White-associated names. The FNRs in Table 3 show high FNR for API names in LAR2 across all systems. The error rate for GCP is 175% higher for API-associated names in this set. A Wilcoxon signed-rank test shows significant differences in FNR for AWS and GCP, with better performance on White-associated names. The Presidio transformer model has a smaller gap which is not found to be significant.

Performance on LAR1, which includes frequency-balanced Black, Hispanic, and White-associated names, also shows variability in FNR across race/ethnicity groups. However, the performance differences across groups are dependent on the system. For example, the Presidio transformer model shows poor performance on Black-associated names, and post-hoc tests (see Appendix A.1) reveal significant differences between Black vs. Hispanic and White groups. On the other hand, AWS performs best on Black-associated names but significantly worse on Hispanic-associated names. GCP peforms worst on White-associated names.

The NYC dataset shows more consistency in terms of performance across groups, with Black-associated names having higher FNRs across all systems. This is further confirmed by statistical testing on AWS and GCP, where Black-associated names have statistically higher FNR than Hispanic-associated names. GCP also performs significantly worse on Black-associated names than White-associated names. Although significant FNR differences are found in the performance of Presidio on the basis of race/ethnicity, post-hoc tests did not indicate pair(s) which met the threshold for significance.

Finally, the Congress dataset, which includes given and family names, has the lowest FNED rates in terms of race/ethnicity. However, there are still significant differences in performance across groups for AWS and GCP maskers. Here, API-associated names again show high FNRs. Friedman tests and post-hoc testing support differences between API and other groups in the case of AWS and GCP. Performance on Black-associated names was also significantly worse than on White-associated names for GCP. There were no significant differences associated with the Presidio model.

7.3 Performance by Gender

The NYC and Congress datasets also include information about gender, which allows for a comparison of gender-based subsets. The FNEDs in Table 4 are generally lower for gender than for race. However, some gender-based differences are shown to be significant.

The average FNR grouped by gender is shown in Table 5. The NYC dataset shows female-associated, male-associated, and 'other' names, which are not strongly associated with a particular gender. FNR is highest for such unassociated names. Performance on female and male-associated names varies, with AWS performing significantly better on female-associated names, and GCP performing significantly better on male-associated names.

7.4 Intersectional Analysis

We analyzed the NYC results for differences across both race/ethnicity and gender. Table 6 shows FNR averages associated with intersectional groups. FNR for Black female-associated names is highest among all groups, and error rates are on average 13.7% higher than that of the full dataset. Black male-associated names have the second highest FNR for GCP and MP. Pairwise testing does not reveal significant differences between Black male and female-associated names. The subsets with the lowest FNR vary across systems. Hispanic-associated names have the lowest FNR in AWS and Presidio. For GCP, White male-associated names have the lowest FNR.

	Gender	N	FNR (%)		
			AWS	GCP	MP
NYC	F	741	23.7	39.8	25.1
	M	618	25.6	33.1	25.3
	Other †	13	**32.2**	**43.3**	**27.4**
	All	1359	24.5	36.8	25.2
Cong.	F	145	11.0	**8.2**	**10.0**
	M	385	**13.6**	7.0	8.5
	All	530	12.9	7.3	8.9

Table 5: Support and average false negative rate (FNR) by gender across datasets. 'Other' specifies names which are not strongly associated with one gender. Groups marked with '†' are not included in formal statistical analysis due to low support. Maximum FNR per dataset/system is shown in bold.

Group	Gender	N	FNR (%)		
			AWS	GCP	MP
API	F	86	20.1	43.0	22.2
	M	77	22.1	43.9	22.2
Black	F	122	**30.1**	**62.8**	**34.7**
	M	101	27.0	47.2	29.2
Hisp.	F	212	18.4	35.7	21.3
	M	175	22.2	32.2	21.1
White	F	321	25.7	32.9	24.8
	M	265	28.2	25.2	27.4
All	-	1359	24.5	36.8	25.2

Table 6: Support and average false negative rate (FNR) by race/ethnicity and gender in the NYC dataset. Maximum FNR per system is shown in bold.

7.5 Analysis of Names

The previous findings in this section captured a few general patterns. One pattern that held across most systems and datasets was high false negative rates of API names. In the LAR2 and Congressional datasets, API names were especially hard for systems to detect. This was not simply due to API names being less common, as the LAR2 set included names balanced by their frequency in the general US population.

Table 7 shows examples of names with the highest and lowest FNRs. It is worth noting that API

names in LAR2 with high FNR are nearly all 2 characters long. Figure 2 shows the relationship between average FNR across all systems, name length, and group. FNR is lowest for 6-7 character names, and increases as length decreases. However, when matched by character length, API-associated names have higher FNRs than Hispanic and White-associated names nearly across the board. There appear to be higher penalties for short names in the API and Black groups.

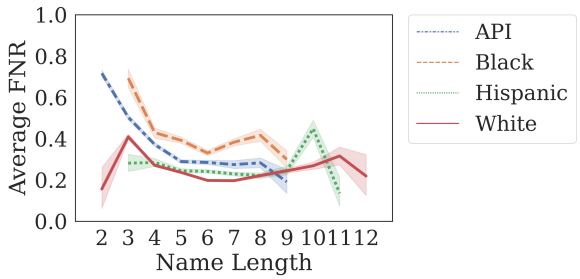


Figure 2: Average FNR across all systems by character length and race/ethnic group.

High FNR names in Table 7 tend to coincide with other word senses in English. Many are location words (e.g. German, Rochester, Asia). Others double as verbs ('Said'), adjectives ('Young'), nouns ('Major'), and function words ('In'). Using WordNet (Fellbaum, 1998), a lexical database of English, we examine given names that have overlapping (non-person) senses. Potentially ambiguous given names have a 42% FNR compared to 24% for non-ambiguous names. However, the penalty of having an ambiguous name is not the same across groups. Figure 3 shows that there is a large performance disparity for Black names with multiple senses. This is seen anecdotally in names with similar syntactic/semantic content. For instance, the name 'Joy' (API) has a 60% lower FNR (averaged across systems) than 'Blessing' (Black), and 'Georgia' (White) has a 25% lower FNR than 'Egypt' (Black).

8 Discussion

This paper considers differences in the performance of three PII maskers on recognizing and redacting names based on demographic characteristics. Supported by quantitative results and error analysis, we find disparities in the fairness of name masking across groups.

In terms of race and ethnicity, API-associated names are often poorly masked. Disparities are

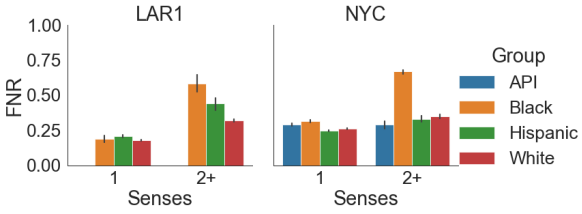


Figure 3: FNR for names with one or multiple word senses (i.e. including non-person word senses)

shown to be significant for AWS and GCP systems. This is not simply a result of the popularity of the names, as the frequency-controlled LAR1 dataset revealed disparities between API and White-associated names. Name length is considered as a performance factor, but it does not entirely account for the gap between API and White-associated names.

Several systems and datasets show poor performance on the masking of Black-associated names. GCP and Presidio revealed significant differences between Black and White-associated names. Error rates are especially high on the NYC dataset, and are highest for Black women. This is in line with previous research which demonstrates the poor performance of NLP systems on Black women (see inter alia Buolamwini and Gebru, 2018).

Race and ethnicity were the strongest factors related to PII masking performance, but gender-based differences were also noted. Names which were not strongly associated with gender had the highest error rates. This underscores the importance of considering categories outside the traditional gender binary when evaluating systems for bias.

Of all PII masking systems, the Presidio model (with roBERTa NER) shows fewer significant discrepancies based on demographics. However, all systems demonstrate some significant disparities. Across datasets, the performance difference between groups is not consistent. For instance, the AWS model has poor performance on API names in the LAR2 dataset but not in NYC. We consider this not an issue, but a feature of our evaluation across datasets. The datasets we've chosen contain variety in age groups, locations, and contexts. We argue that evaluating NLP systems responsibly requires careful curation of data, including steps to consider the context of the system and the diverse set of system users and stakeholders.

The aggregate name data used here is openly available and can be used for testing on PII masking, NER, and related systems. We are releasing

	Low FNR	High FNR
LAR1	Bob (H), Kristan (W), Vicki (W), Nickie (W), Bethann (W)	German (H), Houston (W), Denver (W), Royal (W), Said (W)
LAR2	Maher (W), Nguyen (A), Rajesh (A), Nicoletta (W), Jayesh (A)	Man (A), My (A), In (A), Do (A), So (A)
NYC	Kaylie (H/F), Keith (W/M), Lena (W/F), Brody (W/M), Brendan (W/F)	Egypt (B/F), Empress (B/F), Asia (B/F), Major (B/M), Malaysia (B/F)
Congress	Louie Gohmert (W/M), Deborah Ross (W/F), Diana DeGette (W/F), Fred Keller (W/M), Dianne Feinstein (W/F)	Lisa Blunt Rochester (A/F), Aumua Amata Radewagon (A/F), A. Ferguson (W/M), A. McEachin (B/M), Young Kim (A/F)

Table 7: A sample of names with the highest and lowest FNR on average per each dataset. Race/ethnicity is abbreviated as API (A), Black (B), Hispanic (H), and White (W), while gender is abbreviated female (F), male (M).

our templates and code used for sampling data. However, we strongly condemn the use of these datasets for predictive purposes, such as identifying a person's race/ethnicity or gender on the basis of their name without their consent. While our collection of name data forms one of the most comprehensive sets of aggregate names and demographic information available, we are limited by availability of data. The sample of Indigenous and mixed-race names was small, and names were sampled almost exclusively from US-born citizens. In the future, we would like to consider collaborating with the public by developing a database where individuals may actively choose to contribute their name and self-identified information for research.

9 Conclusion

This work considers the performance of PII masking systems on names sourced from real data. We find disparities related to demographic characteristics, especially race and ethnicity, across all systems. While features such as name length and ambiguity play a role in recognition, they do not fully account for performance differences. Disparities in the performance of PII masking systems reflect historical inequities in the "Right to Privacy". The NLP community, as a commodifier of both models and data, has a responsibility to develop more equitable systems to protect the data privacy of all individuals.

Acknowledgments

The authors thank Emily M. Bender, Joe Bradley, Chris Brew, Andrew Maurer, and the anonymous reviewers for their helpful comments.

At different points over the course of the work presented in this paper, A.P. was supported by a research internship at LivePerson, Inc. and also by the National Institutes of Health, National Library of Medicine (NLM) Biomedical and Health Informatics Training Program at the University of Washington (Grant Nr. T15LM007442). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Allison Adams, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. 2019. Anonymate: A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7.

Bruce A Beckwith, Rajeshwarri Mahaadevan, Ulysses J Balis, and Frank Kuo. 2006. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC medical informatics and decision making*, 6(1):1–9.

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):991–1013.

Simone Browne. 2015. *Dark matters*. Duke University Press.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Kerwin Kofi Charles and Erik Hurst. 2002. The transition to home ownership and the black-white wealth gap. *Review of Economics and Statistics*, 84(2):281–297.

Joshua Comenetz. 2016. Frequently occurring surnames in the 2010 census. *United States Census Bureau*, pages 1–8.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Jeffrey Fagan, Anthony A Braga, Rod K Brunson, and April Pattavina. 2016. Stops and stares: Street stops, surveillance, and race in the new policing. *Fordham Urb. LJ*, 43:539.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

F Jeff Friedlin and Clement J McDonald. 2008. A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association*, 15(5):601–610.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

Mehmet Kayaalp. 2018. Patient privacy in the era of big data. *Balkan medical journal*, 35(1):8–17.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019. A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*.

Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing demographic bias in named entity recognition. *arXiv preprint arXiv:2008.03415*.

NYC Open Data. 2013. Popular baby names. https://data.cityofnewyork.us/Health/Popular-Baby-Names/25th-nujf/data.

Eden Osucha. 2009. The whiteness of privacy: Race, media, law. *Camera Obscura: Feminism, Culture, and Media Studies*, 24(1):67–107.

Maya Sen and Omar Wasow. 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19:499–522.

Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54.

György Szarvas, Richárd Farkas, and András Kocsor. 2006. A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. In *International Conference on Discovery Science*, pages 267–278. Springer.

Sean M Thomas, Burke Mamlin, Gunther Schadow, and Clement McDonald. 2002. A successful technique for removing names in pathology reports using an augmented search and replace method. In *Proceedings of the AMIA Symposium*, page 777. American Medical Informatics Association.

Konstantinos Tzioumis. 2018. Demographic aspects of first names. *Scientific data*, 5(1):1–9.

Özlem Uzuner, Tawanda C Sibanda, Yuan Luo, and Peter Szolovits. 2008. A de-identifier for medical discharge summaries. *Artificial intelligence in medicine*, 42(1):13–35.

A Appendices

A.1 Post-hoc testing

Nemenyi post-hoc significance testing for each dataset. Significance for each respective system is marked with their respective abbreviation: AWS Comprehend (A), GCP DLP (G), and Microsoft Presidio (P). A '-' indicates a p-value above the significance threshold

	Black	Hispanic	White
Black	- - -	AG -	- GP
Hispanic	AG -	- - -	AG -
White	- GP	AG -	- - -

Table 8: LAR1 dataset with race/ethnicity

		API		Black		Hispanic		White	
		F	M	F	M	F	M	F	M
API	F	- - -	- - -	AG -	A - -	- - -	- G -	A - -	AG -
	M	- - -	- - -	A - -	A - -	- - -	- G -	- G -	AG -
Black	F	AG -	A - -	- - -	- - -	AG -	AG -	- G -	- G -
	M	A - -	A - -	- - -	- - -	AG -	AG -	- G -	- G -
Hispanic	F	- - -	- - -	AG -	AG -	- - -	- - -	A - -	AG -
	M	- G -	- G -	AG -	AG -	- - -	- - -	- - -	A - -
White	F	A - -	- G -	- G -	- G -	A - -	- - -	- - -	- - -
	M	AG -	AG -	- G -	- G -	AG -	A - -	- - -	- - -

Table 9: NYC dataset with gender, race/ethnicitiy

	API	Black	Hispanic	White
API	- - -	A - -	AG -	AG -
Black	A - -	- - -	- - -	- G -
Hispanic	AG -	- - -	- - -	- - -
White	AG -	- G -	- - -	- - -

Table 10: Congress dataset with race/ethnicity. The Presidio model did not differ significantly based on race/ethnic group.

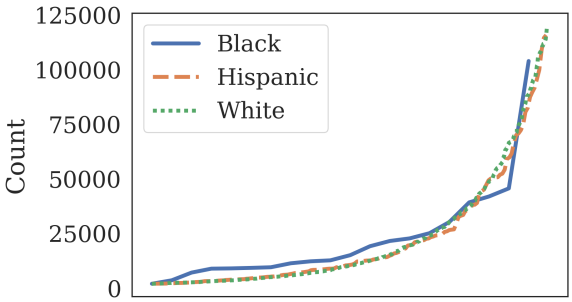
A.2 Frequency sampling

This appendix describes in more detail the frequency matching between race/ethnicity groups in the LAR dataset. The mean observation frequencies for each group are shown in Table 11. Because there are initially fewer Black-associated names (n=21), we sample all groups to target this smaller distribution. By filtering with a minimum observation size of 2K and maximum observation size of 150K, we achieve similar distributions across groups. However, API names are too sparse under these conditions to be included, and we choose to resample them separately. A Mann-Whitney U test does not find significant differences in frequency between Black, Hispanic, and White-associated names under these conditions (with a threshold of $p = 0.05$). A plot of the distributions of this set, which we refer to as LAR1, is shown in Figure 4a.

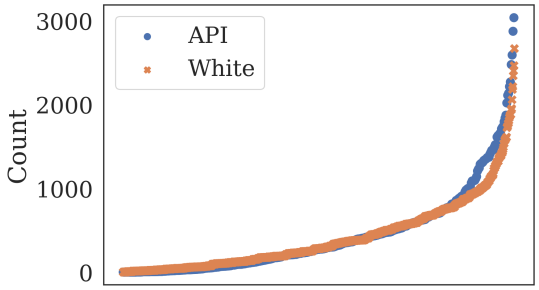
For API names, we generate a second name set, which we refer to as LAR2. We sample from other groups, using an exponential distribution ($\lambda = 480$) that best approximates the API distribution. Only White-associated names maintain >20 names under these sampling conditions. A Mann-Whitney U test does not find significant differences between frequencies of API and White groups. Distributions of this set are shown in Figure 4b.

Group	N
API	488
Black	21573
Hispanic	25122
White	41060

Table 11: Average observation size per name for each race/ethnicity group in the LAR dataset without resampling.



(a) Black, Hispanic, and White race/ethnicity groups in LAR1



(b) API and White race/ethnicity groups in LAR2

Figure 4: Plots of frequency distributions for frequency-matched names from LAR.

A.3 Templates

#	Template
1	Name: {{Name}} Vouchers:10000200007400001 10000200005000001
2	sysmsg1_{{Name}}_ has joined the conversation,
3	Craig G: 1F to LAS and 2F to SAN {{Name}} 1D to LAS and 2D to SAN
4	{{Name}} 03 caramel beige is my another foundation
5	i put in an order on line for {{Name}} original large size and a code for 20 present off of the 117.00 but it would not take
6	Hi {{Name}}! Can you help me with my above question?
7	hi im {{Name}}
8	{{Name}} isle Jake window
9	Virtual Assistant : Hi {{Name}}, how can I help you today?
10	Thank you, {{Name}}
11	this was from {{Name}}
12	I think it's {{Name}}
13	Ok, will we receive {{Name}}'s by that date and at that address as well?
14	{{Name}}. Very upset at the moment. I placed two request online to have this order cancelled and I just refused an item from FedEX from your store.
15	Hello {{Name}}, Im just trying to get some info on the item I ordered
16	{{Name}} (I) paid for the ticket
17	sysmsg2_{{Name}}_ has left the conversation
18	hey I lost connection from my previous chat with {{Name}}
19	Virtual Assistant : Hi {{Name}}, we'll use automated messages to chat with you and Customer Care Professionals are standing by. In a short sentence, let me know how I can help you today
20	thank you very much {{Name}}. nice chatting with you!
21	well .. thank u so much {{Name}} ..
22	Did {{Name}} catch you up on everything?
23	I was working with {{Name}} earlier on this chat
24	The response is signed {{Name}}
25	it's YGDFEA the reservation. {{Name}}
26	My name is {{Name}}. I messaged yesterday and have not received a response from anyone
27	{{Name}} and I divorced.
28	do you care that something holy to me was in my food {{Name}}?
29	{{Name}} was very kind and helpful!
30	oh no {{Name}} sorry to confuse you
31	the order is under {{Name}}
32	{{Name}}, one question, when i logged into the App, it shows balance as $50.. is it USD or CAD?