Detecting intersectionality in NER models: A data-driven approach

Ida Marie S. Lassen
Mina Almasi
Kenneth Enevoldsen
Ross Deans Kristensen-McLachlan
Center for Humanities Computing
Aarhus University, Denmark
idamarie@cas.au.dk, mina.almasi@post.au.dk,
kenneth.enevoldsen@cas.au.dk, rdkm@cc.au.dk

Abstract

The presence of bias is a clear and pressing concern for both engineers and users of language technology. What is less clear is how exactly bias can be measured, so as to rank models relative to the biases they display. Using an innovative experimental method involving data augmentation, we measure the effect of intersectional biases in Danish models used for Named Entity Recognition (NER). We quantify differences in *representational biases*, understood as a systematic difference in error or what is called *error disparity*. Our analysis includes both gender and ethnicity to illustrate the effect of multiple dimensions of bias, as well as experiments which look to move beyond a narrowly binary analysis of gender. We show that all contemporary Danish NER models perform systematically worse on non-binary and minority ethnic names, while not showing significant differences for typically Danish names. Our data augmentation technique can be applied on other languages to test for biases which might be relevant for researchers applying NER models to the study of textual cultural heritage data.

1 Introduction

Issues of bias and discrimination are essential in contemporary Natural Language Processing (NLP). Research has consistently pointed to bias in word embeddings (Kurita et al., 2019; Manzini et al., 2019), and for downstream tasks such as coreference resolution (Zhao et al., 2018), and language generation (Sheng et al., 2021). Several survey papers have also mapped out the landscape of bias research in the field of NLP, showing a lack of clear definitions of bias and normative motivation in NLP bias research (Blodgett et al., 2020); and further emphasising the lack of explicit theorising over the concept of "gender" even when gender biases are the primary concern of a paper (Devinney et al., 2022); and pointing to lack of considerations about the ethical implications of biases in NLP frameworks (Stanczak and Augenstein, 2021).

In this paper, we build on these findings and contribute to ongoing work measuring and quantifying the effects of biases in NLP. We focus on one specific downstream task, namely Named Entity Recognition (NER), and we focus only on the Danish language. We examine error disparities as a function of sensitive features (Borkan et al., 2019; Shah et al., 2020), where earlier work has shown differences across different demographic groups, namely gender and ethnicity (Enevoldsen et al., 2021; Kristensen-McLachlan et al., 2022).

Existing work has highlighted how unintended bias in NLP systems leads to systematic differences in performance for different demographic groups (Borkan et al., 2019; Gaut et al., 2020; Zhao et al., 2018). In response to these results, various frameworks, fairness metrics, and recommendations for the field have been developed to quantify and mitigate bias (Shah et al., 2020; Borkan et al., 2019; Czarnowska et al., 2021; Gaut et al., 2020; Blodgett et al., 2020). Additionally, a growing body of work has demonstrated how *Counterfactual Data Augmentation* (CDA) of training data can be used to mitigate biases in NLP frameworks. This approach has been used for coreference resolution (Zhao et al., 2018), and its applicability has been shown for a broader set of NLP tasks (Lu et al., 2020). We propose another use of data augmentation, namely as a method to *test* the robustness of NLP models and uncover potential social biases in the models.

Informed by intersectional feminism (Crenshaw, 2013), we expand on earlier analysis to investigate the effect of different dimensions of bias and prejudice. The fundamental idea in intersectional feminism relates to how multiple dimensions of inequality result in complex, intersected inequality that cannot be accounted for through an isolated analysis of the single inequalities. For example, minority women might experience other types of discrimination than majority women and still others

Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL2023), pages 116–127
May 5, 2023 ©2023 Association for Computational Linguistics

than those experienced by minority men.

As the discussion and investigation of bias require more than a narrow focus on the overall performance score, adding nuances to bias tests opens up new findings and further reflections. In this paper, we examine how names mainly used by minority communities *and* names used by different genders affect the performance of NER models *together*. Including non-gendered names in our experiment, we furthermore look to challenge the binary understanding of gender dominating the field of bias research.

Our experiments are limited to Danish, a relatively high-resource language from a fairly homogeneous society with a restrictive gendered name law. Our results demonstrate that for contemporary Danish NER, error disparity is *not* evenly distributed across social groups and genders. This result adds significant nuances to the discussion of bias outlined in earlier iterations of this study (Enevoldsen et al., 2021; Kristensen-McLachlan et al., 2022) by highlighting the importance of nuanced perspectives on performance scores (Birhane et al., 2022) and encouraging awareness of who is affected by NLP pipelines.

In drawing attention to differences in performance across sensitive attributes, our focus is on biases as *representational harms* (Crawford, 2017). The harmful aspects derive from the consequences of being excluded from the functionalities of automated systems employed in specific contexts. Communities and individuals who are unrecognised risk falling into the *residual space* of being unseen and treated as irrelevant (Star and Bowker, 2007). In the case of textual cultural heritage, this manifests itself as *archival silence*, the absence of certain voices, stories, and histories (Carter, 2006). We argue that it is vital for those studying textual cultural heritage data with language technology to be able to measure the kinds of bias we outline, in order to avoid reproducing this silence.

2 Bias in NLP

According to one influential definition, bias in computer systems can be defined as systems which 'systematically and unfairly discriminate against certain individuals or groups of individuals in favour of others' (Friedman and Nissenbaum, 1996). This can be further broken down to distinguish between *preexisting biases* with roots in institutions, practices, and attitudes; *technical biases* arising from the resolution of issues in the technical specifications; and *emergent bias* which occurs in a use-context after the implementation of a given system. It has furthermore been suggested to include 'freedom of bias' in the criteria for good computer systems.

Blodgett et al. (2020) provide a survey of bias research in NLP specifically and present a conceptual framework to characterize and compare biases. Drawing on earlier work (Crawford, 2017), they distinguish between *allocation bias* and *representational bias*. The former is a difference in the allocation of resources and opportunities; while the latter is differences in representations, such as stereotyping and negative generalisation of social groups. Representational bias furthermore includes differences in system performance, such as how well an automated system performs for different demographic groups.

We focus on *representational bias* as differences in system performance, measured as differences in error on a particular task. Crawford (2017) emphasise representational bias as *harmful* in itself, mirroring the ideas of an *emergent bias* in Friedman and Nissenbaum (1996). Further bias emerges when systems whose performance differs systematically across different demographic groups are implemented.

In social science and humanities, researchers who apply NLP tools in their work need to consider such performance differences when deciding which framework to use. For example, a researcher working in the field of gender history might need their models to be particularly robust with respect to gender; a scholar of social media might have a specific reason to require that their model is particularly robust to different ethnicities represented in their data. For those who work with cultural heritage data, there may therefore need to be a necessary trade-off between the overall accuracy of a particular framework and the bias that it exhibits relative to different groups.

In the following section, we outline how existing societal biases in Denmark make it crucial to test NLP frameworks for technical biases.

2.1 Intersections of discrimination

Injustices encountered by social groups can rarely be accounted for through a single variable (such as either gender or race) but interacts with other systems of oppression (such as race, age, class,

$$\begin{array}{rcc} & \textit{majority} & \textit{minority} \\ \textit{women} & \mathcal{A} & \mathcal{B} \\ \textit{men} & \mathcal{C} & \mathcal{D} \end{array}$$

Figure 1: The intersectional subgroups $\mathcal{A}$ = majority women, $\mathcal{B}$ = minority women, $\mathcal{C}$ = majority men, $\mathcal{D}$ = minority men, are defined by combinations of senstive attributes – in this case gender and ethnicity (Subramanian et al., 2021).

disabilities, education level, etc.). This has been termed *intersectionality* as different dimensions of oppression intersect and affect the encountered injustice (Crenshaw, 2013).

Nevertheless, most research on bias in machine learning - and in NLP specifically - focuses on a single dimension of discrimination, most often either race (Field et al., 2021; Manzini et al., 2019) or gender (Kurita et al., 2019; Basta et al., 2019). If multiple bias markers are examined, the combined effect is often left out of the picture (Garg et al., 2018; Czarnowska et al., 2021; Nadeem et al., 2021). However, recent work has shed light upon intersectional biases in NLP. In particular, Lalor et al. (2022) benchmark multiple NLP models on fairness and predictive performance across various NLP tasks. They deploy multiple demographic dimensions and evaluate various downstream NLP tasks for *allocation biases*. Furthermore, Subramanian et al. (2021) evaluate different debiasing techniques and suggest a post-hoc debiasing method particularly useful for intersectional biases. In a more analytical line of work, Herbelot et al. (2012) provide a quantitative analysis of concepts from gender studies and presents a methodological approach to the investigation of intersectional bias at the level of word representations.

In this paper, we examine *representational bias* in named entity recognition in Danish NLP frameworks. We define bias as a difference in system performance measured by error rate as a function of sensitive features – gender and ethnicity. To test the error disparities for NER across different demographic groups, we divide our data set into subgroups functioning as proxies for the demographic subgroups in question. To do so, we use gender-divided name lists with minority and majority names, which allow us to conduct an intersectional analysis of the effect of different oppressive dimensions. We furthermore include unisex names in our experiments in an attempt to move beyond binary conceptions of gender.

2.2 Muslim names as a proxy for ethnic minority

With names come strong connotations to both ethnicity and religion, and a name often reveals group affiliation for individuals (Khosravi, 2012). In Denmark, the largest immigrant community has members descended from Middle Eastern and Muslim countries (Statistics Denmark, 2022). Research has pointed out how people in this group experience various types of discrimination spanning from harsh rhetoric in political discourse over ministerial administration (Vinding, 2020) to hate crimes (Mannov, 2021) and exclusion of labour market (Dahl and Krog, 2018).

Given the sociological evidence, it is clearly worth considering the impact of machine learning technologies for a large part of the Danish population who is vulnerable to discrimination (Jørgensen, 2023, Ranchordás and Scarcella, 2021). A list of Muslim first names used in Denmark was retrieved from Meldgaard (2005), which presented the names of Muslim origin used in Denmark in 2005, together with an explanation of the meaning of each name. As most immigrants in Denmark come from predominantly Muslim countries, we apply this list of names as a proxy for minority ethnicity. The list is furthermore divided into women's and men's names.

Of course, not all minority people in Denmark will be represented on this list of names, which represents a known limitation of our work. Instead, we infer only ethnicity on a group level, and as research has shown that Middle Eastern immigrants are being subjected to discrimination on the basis of their names (Dahl and Krog, 2018), we argue that testing performance for this group is a necessary step for quantifying bias in NLP frameworks.

A name of Muslim origin might, however, not be the only source of discrimination. Experiments in which fictitious job applications were randomly assigned either a Danish or Middle Eastern-sounding name and sent to actual job openings showed that minority men are consistently subject to a much larger degree of discrimination than minority women (Dahl and Krog, 2018). Similarly, experiments on commercial automated facial analysis systems for gender classification showed that women with darker skin are the most misclassified group (Buolamwini and Gebru, 2018). Hence, the protected and privileged group might vary across contexts, and to examine the intersection between ethnicity and gender discrimination, a proxy for gender is needed.

2.3 Names as a proxy for gender

Denmark has a high level of formal equality, with anti-discrimination laws ensuring constitutional equality and discrimination protection. However, structural oppression still exists and can be shown in studies on the gender pay gap (Gallen et al., 2019) as well as in statistics on violence against women (European Union Agency For Fundamental Rights, 2014). The work by Dahl and Krog (2018) furthermore showed that in a labour market context, women were subject to discrimination except in the women-dominated fields, where men experienced a slightly lower call-back rate.

Using a gendered name list as a proxy has advantages and disadvantages. On the one hand, demarcating our results on proxies for gender *and* ethnicity allows us to conduct an intersectional analysis of the relative effect of gender and ethnicity on the error disparities. Denmark has strict name laws relative to many other European countries, restricting which names a person can be assigned according to their gender (something which has been actively criticised by citizen activist groups[1]). Hence we are not only relying on a majority count of the usage of a name but on a legal context determining the 'gender' of a name – highly dominated by a binary understanding of gender.

On the other hand, the disadvantages of augmenting on gendered name lists are the risk of reinforcing a folk conception of gender (Keyes, 2018), where gender is understood as binary and static, and ruling out other gender identities (Dev et al., 2021). Danish names are neither inherently nor definitively gendered, and the implementation of laws restricting the choice of name based on sex assigned at birth emphasises how ideology is present both in Danish name laws and in the language in general (Blodgett et al., 2020).

Instead of a bio-essential binary understanding, gender can be conceptualised as both performative and constituted by discursive practices (Butler, 2006). With such an understanding of gender, biological sex and cultural gender are separated, and neither can be inferred from a name or physiological appearance. Introducing ourselves with certain names and pronouns can be one way of performing a gender but is not the only way. It may, therefore, still be problematic to use gendered name lists to infer the gender of an individual. However, as mentioned above, we only infer at the group level to assess the potential biases for different demographic groups when subjected to NLP frameworks. Furthermore, we do not link names to pronounces and do not draw conclusions about individual *gender identity*.

In an attempt to go beyond a solely binary understanding of gender, we include unisex names which are culturally understood as being used by both men and women. However, it should go without saying that non-binary people do not specifically use these names, and it might be an insufficient way of challenging the binary concept of gender.

We do not claim that these proxies for either gender or ethnicity are perfect. However, as we do not infer values of sensitive attributes at the level of individuals but examine structural differences at the group level, we find these proxies highly productive for examining differences in system performance for different demographic groups and to expand earlier analysis by considering the intersection of oppressive dimensions.

Given these qualifications, our experiment in data augmentation is motivated by the following research questions:

- **RQ1** Does system performance differ across the subgroups shown in Figure 1?
- **RQ2** Does system performance differ for unisex names compared to majority names?
- **RQ3** Does system performance for the different groups differ across the selected NLP frameworks?

In order to answer these questions, we test the system performance on all known systems for performing Danish NER.

3 Method

We define bias as the systematic difference in error, *error disparity*, as a function of a given sensitive feature (Shah et al., 2020). We deploy *Counterfactual Data Augmentation* (CDA) (Lu et al., 2020), not as a way of debiasing the framework, but as a test method for examining *error disparity* across different sensitive features. In other words, bias in the model is measured through the difference

[1] See Ligebehandling for alle (2021) for citizen proposal for abolishing of the gender-separated name lists including critique and explanations (is available in Danish).

in performance accuracy when data is augmented with different gender and ethnicity features.

In Enevoldsen et al. (2021), a range of contemporary Danish NLP frameworks was subjected to a series of data augmentation strategies to test their robustness during training. These augmentations included random keystroke augmentation to simulate spelling errors; and spelling variations specific to the Danish language. Additionally, among the augmentation strategies were the following name augmentations:

1. Substitute all names (PER entities) with randomly sampled majority names, respecting first and last names.
2. Substitute all names with randomly sampled minority names (Meldgaard, 2005), respecting first and last names.
3. Substitute all names with sampled majority men's names, respecting first and last names.
4. Substitute all names with sampled majority women's names, respecting first and last names.

These augmentations specifically tested the robustness of named entity recognition in each Danish NLP framework, given data augmented relative to gender and ethnicity. If a framework performed just as well (or better) with these augmentations as without, this was interpreted as an indicator of robustness. Conversely, if a framework performed worse, our approach makes it possible to quantify exactly where the model is failing and, hence, where potential biases reside.

We expand on this analysis by testing the disparities in performance across different dimensions of sensitive attributes, namely gender and ethnicity. This is done by dividing minority names into gender. Instead of relying on a solely binary conception of gender, we furthermore test the robustness of named entity recognition in each Danish NLP framework for names on the unisex name list.

Hence, adding to the above list:

5. Substitute all names with sampled minority women's names, respecting first and last names.
6. Substitute all names with sampled minority men's names, respecting first and last names.
7. Substitute all names with sampled unisex names, respecting first and last names.

3.1 Danish NLP frameworks

We have attempted in this experiment to draw on all existing frameworks which can be used to perform NER on Danish language data. Each framework uses different architectures and training data.

spaCy uses pre-trained word-embedding initialised using a tok2vec component[2]. For the purposes of this experiment, we have not included spaCy's Transformer-based model, *da_core_news_trf*, since it corresponds to the DaCy-medium outlined below.

DaCy (Enevoldsen et al., 2021) is a unified state-of-the-art framework for Danish NLP built on spaCy. DaCy-small is based on a Danish Electra (14M parameters); DaCy-medium is based on the Danish BERT (110M parameters)[3]; and DaCy-large is based on the multilingual XLM-Roberta (550M parameters).

ScandiNER[4] is a model trained for NER across many Scandinavian languages including Danish. The model itself is a finetuned BERT-base model trained on the digitised collections of the Norwegian national library[5]. While explicitly referred to as a Norwegian model, it has been trained on a wide range of data and has proven to be highly performant on Danish text data [6].

Flair (Akbik et al., 2019) is a BiLSTM-based model which has demonstrated high levels of performance on Danish as well as similar languages, such as English and German. BiLSTM models tend to be computationally more expensive to train than Transformers due to their use of recurrence. However, BiLSTM models like Flair continue to be popular and are hence included in our experiment.

Polyglot employs a static word embedding model using word embeddings trained on Wikipedia (Al-Rfou' et al., 2013). While not as widely used as it once was, we have included this model to illustrate differences in performance between older models and more state-of-the-art Transformer-based models.

Many of these models are built on top of BERT-style architectures. In the case of English, models from this family have been shown to encode spe-

[2] https://explosion.ai/blog/deep-learning-formula-nlp

[3] https://huggingface.co/Maltehb/danish-bert-botxo

[4] https://huggingface.co/saattrupdan/nbailab-base-ner-scandi

[5] https://huggingface.co/NbAiLab/nb-bert-base

[6] https://scandeval.github.io/

Data set overview	All	Filtered
Nr. of unisex first names	500	500
Nr. of majority first names	1,000	943
women's names	500	485
men's names	500	458
Nr. of majority last names	500	500
Nr. of minority first names	1,134	1,121
women's names	452	443
men's names	625	621
Nr. of minority last names	526	526

Table 1: The number of names used in the data augmentation: The left column is the number of names; 500 majority names for men, women, and unisex are chosen to match the number of minority names. The right columns show the number after the overlap between majority and minority lists is filtered away. The number of minority women's and men's names do not amount to the total number of minority names due to an overlap of names, which is not filtered out.

cific biases across multiple axes of discrimination (Bender et al., 2021). It has also been demonstrated that BERT-style models have a tendency to learn stereotypical representations (Kurita et al., 2019). Previous work has shown that all Danish models exhibit statistical significant bias in terms of ethnicity, while only Polyglot shows a gender bias (Enevoldsen et al., 2021). As such, we expect to see similar results when testing Danish NER models, with poorer performance for the subgroups marginalised among more than one dimension.

All models are fine-tuned on the DaNE dataset (Hvingelby et al., 2020) with the exception of Polyglot, which is trained using the Wikipedia data.

3.2 Data

As described in Section 2 the list of minority names is retrieved from Meldgaard (2005) containing $\sim 1,000$ names. For minority last names, a list of Muslim last names are retrieved from FamilyEducation[7]. The majority and last names lists are retrieved from Statistics Denmark[8], filtered on the 500 most used names for men, women, and last names to approximately match the number of minority names. Finally, the list of unisex names is retrieved from The Agency of Family Law[9] we have filtered on the 500 most popular unisex names according to the data from Statistics Denmark.

As the list of Danish names consists of popular names in Denmark, there is an overlap of 75 names also classified as minority names according to the list from Meldgaard (2005). To report the true effect of the minority names, we have filtered out those such that they only appear in the list of minority names - resulting in 458 majority men's names and 485 majority women's names.

For example, as Mohammed is a common name in Denmark with Islamic origin, it occurs in both the majority and minority name lists. However, Mohammed is most likely a name being subjected to discrimination in line with the work by Dahl and Krog (2018). Therefore, we have filtered it out to only occur on the list of minority names. However, we found some names impossible to classify as *either* majority or minority names, and we included them in both lists. This includes names like Sara, Sarah, Laila, and Ben. A similar sorting of the overlap between the gendered name lists and the unisex name lists is not meaningful, as it is the very definition of the unisex names that they can be used by all genders. Table 1 provides an overview of the number of names for each category[10].

The experimental pipeline is set up as follows. For each sentence in the DaNE dataset, we augment the dataset by replacing each "PERSON" entity with a name randomly sampled from one of the given lists. To avoid nonsensical sentences, we ensure that within one document, a specific name is always replaced by the same name. Following this, the NER performance for all models is tested on the augmented data, estimated by calculating F1 scores across all tags. As the random choice of name influences the performance, we repeat this process 20 times for each model to estimate a mean F1 score. Finally, we used a t-test to compare whether the F1 scores obtained on the augmented data varied significantly from the baseline. For the baseline, we used the majority names for both genders (see Table 2. As we perform multiple comparisons, we make sure to adjust the p-values using a Bonferroni correction.

The name augmentation was performed using Augmenty (Enevoldsen, 2022) and the model evaluation was performed using DaCy framework. All code is publicly available and open source, shared

[7] https://www.familyeducation.com/baby-names/surname/origin/muslim

[8] https://www.dst.dk/da/Statistik/emner/borgere/navne/navne-i-hele-befolkningen

[9] https://familieretshuset.dk/navne/navne/godkendte-fornavne

[10] See https://github.com/centre-for-humanities-computing/Danish-NER-bias/tree/main/name_lists for complete name lists

using an Apache 2.0 license[11].

4 Results

In Table 2, we see the results of the name augmentation experiments. We see that larger, transformer-based models consistently outperform other models on NER tasks. These results underline three well-known trends in deep learning and NLP: 1) larger models tend to perform better than smaller models; 2) higher quality pre-training data leads to better models; and 3) multilingual models perform competitively with monolingual models (Brown et al., 2020; Raffel et al., 2020; Xue et al., 2021).

More pertinently, our results show that the NER performance of every model is affected by the data augmentations. It is immediately apparent, though, that not all models are affected equally, and not all augmentations cause pronounced effects. Our results seem to demonstrate that Danish language models are relatively more robust to the impact of randomly changing women and men's names *at the majority level*. However, this is not the case for unisex names, where our results show that *all* Danish NLP models are significantly worse at recognising these compared to gender-conforming names.

Similarly, randomly replacing names with minority names results in significantly worse performance for all models. This suggests that Danish NLP models contain a greater relative bias regarding ethnicity than the binary gender division, emphasised by the results showing that all models performed consistently better for majority women's names than for minority men's names. For **ScandiNER**, all **DaCy** models, **DaNLP BERT**, **Flair**, and **NERDA**, the performance for minority women's names and minority men's names are similar - but still significantly lower than names from 'majority all'. For **Polyglot** and the **spaCy** models, the performance for minority women is worse than those for minority men. Especially interesting are the results from **DaCy Large**, where there is no apparent bias for minority names if intersectionality is left out of the picture. However, a bias towards minority women is shown when minority names are divided into men's and women's names. **ScandiNER** performs overall best of all models, and even though it shows bias towards 'minority all' and 'minority men', it still outperforms **DaCy Large**, which do not show the same bias in error rate for these groups.

One could argue that for the best performing models, **ScandiNER**, and the **DaCy** models, the differences in F1 scores are overall negligible. However, as small differences accumulate when used on large corpora we argue that even seemingly small differences (which are statistically significant) should be taken into consideration in an NLP pipeline. This becomes more pronounced when one considers the *increase in error rate.* The best performing model is **ScandiNER** shows a 7% increase in error from 'majority all' to 'minority women'. Similarly, for **DaCy medium** this amounts to an increase in error rate of 17%. For the poorest performing **Polyglot** model, we calculate a 72% increase in error rate.

Hence, according to Figure 1, we conclude that Danish NLP frameworks perform best for subgroups $\mathcal{A}$ and $\mathcal{C}$ (majority people). On the other hand, the models perform significantly worse for subgroups $\mathcal{B}$ and $\mathcal{D}$ (minority people), and some models are worst for subgroup $\mathcal{B}$ (minority women) specifically. Adding unisex names and the gendered demarcation of the minority lists (see 5-7 in the list in Section 3) to our tests shows that the error disparity is *not* evenly distributed across the social groups in Figure 1. These results open up the narrow focus on the overall performance scores and are significant contributions to the examination of bias started in earlier iterations of this study (Enevoldsen et al., 2021; Kristensen-McLachlan et al., 2022).

5 Discussion

Much of the work on representational bias focus on system performance and the concrete impact on individuals and groups as a result of biased models. However, we argue that similar considerations should underlie research applications of NLP, such as the use of language technology to study cultural heritage data. By ignoring the disparate performance of NLP frameworks on downstream tasks, we risk overlooking the testimony of marginalised voices in our corpora and archives.

Previous work has outlined how, in classification systems, residual categories are those that are left out when categories are established (Star and Bowker, 2007; Scheuerman et al., 2019). By not complying with the agreed-upon categories, the 'other' fall between the cracks of the categorisation

[11] See `https://github.com/centre-for-humanities-computing/Danish-NER-bias` for code for the experimental pipeline

	All		Men		Women		Unisex
Model	Majority	Minority	Majority	Minority	Majority	Minority	Majority
ScandiNER	89.1(0.4)	**88.3(0.6)***	89.0(0.5)	**88.4(0.4)***	89.0(0.4)	**88.3(0.6)***	**88.6(0.4)***
DaCy large	86.7(0.5)	86.4(0.6)	86.6(0.4)	86.3(0.4)	86.5(0.3)	**86.0(0.6)***	**86.2(0.5)***
DaCy medium	79.9(0.6)	**77.0(0.8)***	79.6(0.5)	**76.4(1.1)***	79.9(0.5)	**76.6(0.7)***	**78.2(0.8)***
DaCy small	77.8(1.0)	**74.8(1.0)***	77.8(0.8)	**74.6(1.2)***	77.6(0.8)	**74.9(1.1)***	**76.0(1.0)***
DaNLP BERT	83.4(0.5)	**81.1(1.0)***	83.4(0.4)	**81.1(0.7)***	83.6(0.5)	**80.8(0.9)***	**81.9(0.8)***
Flair	81.8(0.4)	**79.9(0.8)***	82.1(0.5)	**79.9(0.8)***	81.6(0.4)	**80.0(0.7)***	**79.8(0.8)***
NERDA	80.6(0.8)	**78.5(1.1)***	81.1(0.8)	**78.7(0.8)***	80.8(0.4)	**78.5(0.7)***	**79.8(0.9)***
SpaCy large	79.0(0.5)	**68.7(1.3)***	79.3(0.6)	**71.2(0.9)***	78.8(0.6)	**66.4(1.7)***	**75.8(0.8)***
SpaCy medium	78.2(0.8)	**64.6(1.4)***	78.7(0.5)	**66.7(1.8)***	78.3(0.5)	**61.0(1.2)***	**71.9(1.3)***
SpaCy small	64.8(0.7)	**57.5(1.4)***	64.6(1.3)	**57.5(1.5)***	65.1(1.4)	**56.3(1.5)***	**61.5(1.4)***
Polyglot	64.9(0.9)	**41.7(1.3)***	**66.1(0.7)***	**42.1(1.2)***	**63.3(1.4)***	**39.5(1.0)***	**57.4(1.5)***

Table 2: Named Entity Recognition (NER) performance of Danish Natural Language Processing (NLP) pipelines reported as average F1 scores excluding the MISC category on the test set. The column 'Majority All' names is considered the baseline for the augmentation of minority, women's, men's and unisex names. Bold and * denotes that the result is significantly different from the baseline using a significance threshold of 0.05 with Bonferroni correction for multiple comparisons. Values in parentheses denote the standard deviation.

schema. This can happen if the object is too complicated to classify in the often taken-for-granted categories or if the residual is unknown to the system. Falling into a residual space can result in people's experience being disregarded or overlooked, consciously or otherwise. In the context of named entity recognition, the classification performed is either recognised or unrecognised, and we argue that people whose names are unrecognised by automated systems reside in the *residual* spaces.

Our results show that, for contemporary Danish NER, there are differences in performance along different demographic lines – differences that may not have been obvious without testing performance for the different subgroups. This, first and foremost, highlights the importance of challenging the narrow focus on overall performance score (Birhane et al., 2022) and sheds light upon the existence of diversity in who is affected. Furthermore, these results also show a difference in the risk of residing into the residual space and potentially being disregarded and mistreated. The technical biases in these NLP frameworks risk reinforcing the existing structural biases if put into use. Therefore, we recommend NLP practitioners to take accountability (Buolamwini and Gebru, 2018) and consider these subgroup-specific performance results. The responsibility of measuring and mitigating such biases should be placed on those developing and implementing the tools – not on the marginalised group who are unfairly treated by the systems (Bender et al., 2021).

In this work, we defined bias as the difference in error rate across different demographic subgroups. This bias is only tested for one specific task. For our data augmentation, we used the DaNE corpus, which consists of a diverse set of written and spoken Danish from 1983–1992. However, minority names might occur more frequently in contexts which differ substantially from this corpus. If this is the case, our reported performances might vary according to how well Danish NLP frameworks perform on NER for minority names 'in the wild'. Hence, assessing the potential bias towards minority people might be even more complex.

A similar issue arises when approximating ethnicity for social groups through the use of name lists. This approach leaves out minority people who take names typical for the majority group. However, when it comes to the performance of NER tools, minority people with majority names are not at the same risk of being unfairly treated by NER tools as people with minority names. The reverse is also the case: a person from the ethnic majority with a name typical for the minority is at greater risk of not being recognised by NER tools than people with majority names. Nevertheless, this is not central to our analysis, insofar as we are only inferring at *group level* when examining the distribution of error rates across different social groups.

Further complexities in the use of names are the effect of rare names. For unisex names, we included the 500 most used names, which are approved unisex names in Denmark. In this list, there are names that are common gendered names, such as 'Anne', which in Denmark is primarily used by women. If we filtered out common and primarily gendered names, the performance might be even poorer, but then it might be an effect of rare names rather than unisex names.

Nevertheless, this paper presents an innovative experimental method which adds nuanced perspectives to the overall performance evaluation for these models. Based on data augmentation and the use of name lists as proxies for multiple dimensions of inequality, the method allows for an intersectional analysis of biases in Danish NLP models used for named entity recognition. Such findings are important to incorporate into scholarly pipelines in order to avoid enforcing *archival silence*.

6 Conclusions

In this paper, we have shown the importance of intersectional analysis of biases in Natural Language Processing (NLP) frameworks by testing Danish NLP frameworks' robustness to data augmentation in Named Entity Recognition (NER).

By augmenting test data on gender-divided name lists for both majority and minority names, we have shown that Danish NLP frameworks are relatively robust to the impact of women's and men's names *at the majority level*. However, *all* Danish NLP models are significantly worse at recognising unisex names compared to gender-conforming names. Furthermore, minority names cause significantly worse performance for all models. This suggests that Danish NLP models contain a greater relative bias regarding ethnicity than the binary gender division.

In the context of textual cultural heritage data, researchers regularly and increasingly incorporate language technology into their scholarly workflow. The most appropriate tool for a given task such as NER is usually chosen based on some pre-calculated metric score for how the technology performs for that task. However, based on the results presented here, we argue that a raw performance measure should not be the only criterion for deciding which NLP model to use. Instead, we emphasise that, in the case of textual cultural heritage data, *accuracy is not all you need*. We encourage researchers to take these sub-group-specific performance measures into account when setting up their research pipeline.

7 Limitations

The current study has some limitations. Firstly, our minority-majority categorisation is rather restricted, and a large group of the population will not be represented in this division insofar as we only include names of primarily Muslim backgrounds, excluding other minority ethnic communities in Denmark. In addition, our approach of manually sorting names which occur in both the majority and minority name lists is potentially problematic as our sorting is based on our (perhaps stereotyped) ideas of these names and not on any shared methodology.

Furthermore, gendered name lists corresponding to Danish name laws rely on, and so reinforce, a binary understanding of gender. We argue that these demarcations in our data are useful for understanding the societal biases which can be embedded in NLP frameworks but are not comprehensive.

Further work is needed to conclude the overall bias level of Danish NLP frameworks. In particular, bias tests for coreference resolution and word embeddings should be conducted. In addition, our work presented experimental results for a single, comparatively small Indo-European language. We would like to see similar experiments conducted on different languages, given an appropriate change of experimental conditions, to see if results are reproduced in different cultural contexts.

8 Ethics Statement

In this work, we have actively engaged with the fact that the actions of machine learning and NLP engineering can change the world and affect both society and individuals. The use of computer technologies may produce new or reproduce existing discrimination, and we, therefore, strive towards being as inclusive as possible. Not only do we wish to draw attention to social biases inherent in contemporary Danish language technology but we hope that our work can be used directly by other researchers when deciding on tools usages in their scholarly pipeline, particularly for those working with cultural heritage data.

9 Online Resources

See `https://github.com/centre-for-humanities-computing/Danish-NER-bias` for code for the experimental pipeline and complete name lists used in data augmentation.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59. Association for Computational Linguistics.

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623. Association for Computing Machinery.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Judith Butler. 2006. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge.

Rodney G.S. Carter. 2006. Of things said and unsaid: Power, archival silences, and power in silence. *Archivaria*, 61:215–233.

Kate Crawford. 2017. The trouble with bias - nips 2017 keynote - kate crawford #nips2017. [Online; accessed 18-February-2023], published by *The Artificial Intelligence Channel*.

Kimberlé Williams Crenshaw. 2013. Mapping the margins: Intersectionality, identity politics, and violence against women of color. In *The public nature of private violence*, pages 93–118. Routledge.

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.

Malte Dahl and Niels Krog. 2018. Experimental evidence of discrimination in the labour market: intersections between ethnicity, gender, and socio-economic status. *European Sociological Review*, 34(4):402–417.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2083–2102.

Kenneth Enevoldsen. 2022. Augmenty: The cherry on top of your NLP pipeline.

Kenneth Enevoldsen, Lasse Hansen, and Kristoffer Nielbo. 2021. Dacy: A unified framework for danish nlp. In *CEUR Workshop Proceedings*, pages 206–216, Amsterdam, The Netherlands. CHR 2021: Computational Humanities Research Conference.

European Union Agency For Fundamental Rights. 2014. Violence against women: an eu-wide survey. Technical report, European Union Agency For Fundamental Rights.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347.

Yana Gallen, Rune V Lesner, and Rune Vejlin. 2019. The labor market gender gap in denmark: Sorting out the past 30 years. *Labour Economics*, 56:58–67.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online. Association for Computational Linguistics.

Aurélie Herbelot, Eva von Redecker, and Johanna Müller. 2012. Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 45–54, Avignon, France. Association for Computational Linguistics.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A named entity resource for danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, LREC '20, pages 4597–4604.

Rikke Frank Jørgensen. 2023. Data and rights in the digital welfare state: the case of denmark. *Information, Communication & Society*, 26(1):123–138.

Os Keyes. 2018. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22.

Shahram Khosravi. 2012. White masks/muslim names: immigrants and name-changing in sweden. *Race & class*, 53(3):65–80.

Ross Deans Kristensen-McLachlan, Ida Marie S. Lassen, Kenneth Enevoldsen, Lasse Hansen, and Kristoffer Laigaard Nielbo. 2022. Accuracy is not all you need. In *DH2022 Tokyo Book of Abstracts*, pages 281–284.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

John P Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in nlp. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609.

Ligebehandling for alle. 2021. Ligebehandling for alle: Afskaf de kønsopdelte navnelister. [Online; accessed 18-February-2023], published by borgerforslag.dk.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202.

Jonas Mannov. 2021. Fakta om hadforbrydelser. [Online; accessed 31-January-2023], published by *The Danish Crime Prevention Council*.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Eva Villarsen Meldgaard. 2005. Muslimske fornavne i danmark. Publisher: Københavns Universitet.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Sofia Ranchordás and Luisa Scarcella. 2021. Automated government for vulnerable citizens: intermediating rights. *Wm. & Mary Bill Rts. J.*, 30:373.

Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Susan Leigh Star and Geoffrey C Bowker. 2007. Enacting silence: Residual categories as a challenge for ethics, information systems, and communication. *Ethics and Information Technology*, 9:273–280.

Statistics Denmark. 2022. Fakta om indvandrere og efterkommere i danmark. [Online; accessed 31-January-2023].

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2498, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Niels Valdemar Vinding. 2020. Discrimination of muslims in denmark. In *State, Religion and Muslims*, pages 144–196. Brill.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.