# Occupational Biases in Norwegian and Multilingual Language Models

**Samia Touileb**
University of Bergen
samia.touileb@uib.no

**Lilja Øvrelid**
University of Oslo
liljao@uio.no

**Erik Velldal**
University of Oslo
erikve@uio.no

## Abstract

In this paper we explore how a demographic distribution of occupations, along gender dimensions, is reflected in pre-trained language models. We give a descriptive assessment of the distribution of occupations, and investigate to what extent these are reflected in four Norwegian and two multilingual models. To this end, we introduce a set of simple bias probes, and perform five different tasks combining gendered pronouns, first names, and a set of occupations from the Norwegian statistics bureau. We show that language specific models obtain more accurate results, and are much closer to the real-world distribution of clearly gendered occupations. However, we see that none of the models have correct representations of the occupations that are demographically balanced between genders. We also discuss the importance of the training data on which the models were trained on, and argue that template-based bias probes can sometimes be fragile, and a simple alteration in a template can change a model's behavior.

## 1 Introduction

Measuring the presence of stereotypical representations of occupations in pre-trained language models has been an important effort in combating and reducing possible representational harms (Blodgett et al., 2020). However, and as pointed out by Blodgett (2021), most of the current work is motivated by an idealised vision of the world where occupations should not be correlated with genders, and where the expectations are that models should not be stereotypical when *e.g.,* predicting female or male pronouns in relation to occupations. The idea that we are all equal is an important factor in our quest of reaching fair and less biased models, and reflect our normative judgments.

While this is true for most stereotypes, it might not directly apply to occupations. With a descriptive and realistic view of the society, there clearly exists gender disparities in occupations. This is inherently tied to many societal constructs and cultural backgrounds, and are a reality for many occupations. Also pointed out by Blodgett et al. (2020), the importance of the connection between language and social hierarchies, has not been considered in most previous work on bias in NLP. It is a reality that most Norwegian nurses are females. Having a model reflecting this reality might not be problematic *per se*, but using this disparity to for example systematically reject male applicants to a nurse position is a very harmful effect.

In this paper, we investigate how the real-world Norwegian demographic distribution of occupations, along the two gender dimensions male versus female, is reflected in large transformer-based pre-trained language models. We give a descriptive assessment of the distribution of occupations, and investigate to what extent these are reflected in four pre-trained Norwegian and two multilingual models. More precisely, we focus on the following research questions:

- To what extent are demographic distributions of genders and occupations represented in pre-trained language models?
- How are demographically clearly gender-correlated vs. gender-balanced occupations represented in pre-trained language models?

To address these questions, we investigate the correlations of occupations with Norwegian gendered pronouns and names. We analyse five template-based tasks, and compare the outputs of the models to real-world Norwegian demographic distributions of occupations by genders.

After first providing a bias statement in Section 2, we give an overview of previous relevant work in Section 3. Section 4 describes our experimental setup, and outlines our template-based tasks. We present and discuss our main results and findings in

*Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200 - 211
July 15, 2022 ©2022 Association for Computational Linguistics

Section 5 and 6. We conclude with a summary of our work, and discuss our future plans in Section 7.

## 2 Bias statement

We follow the bias definition of Friedman and Nissenbaum (1996), where bias is defined as the cases where automated systems exhibit a systematic discrimination against, and unfairly process, a certain group of individuals. In our case, we see this as reflected in large pre-trained language models and how they can contain skewed gendered representations that can be systematically unfair if this bias is not uncovered and properly taken into account in downstream applications. Another definition of bias that we rely on is that of Shah et al. (2020), where bias is defined as the discrepancy between the distribution of predicted and ideal outcomes of a model.

We focus on the associations between gendered (female and male) pronouns/names and professional occupations. We investigate to what degree pre-trained language models systematically associate specific genders with given occupations. However, we explore this from the perspective of a descriptive assessment: Instead of expecting the system to treat genders equally, we compare how these gender–occupation representations reflect the actual and current Norwegian demographics. This will in no way reduce the representational harms of stereotypical female and male occupations, that could both be propagated and exaggerated by downstream tasks, but would rather shed light on which occupations are falsely represented by such models. Moreover, our work will provide knowledge about the biases contained in these models that may be important to take into account when choosing a model for a specific application.

Arguably, a limitation of our work is that we are only able to evaluate correlations between occupations and the binary gender categories male/female, although we acknowledge the fact that gender as an identity spans a wider spectrum than this.

## 3 Background and related work

Training data in NLP tasks may contain various types of bias that can be inherited by the models we train (Hovy and Prabhumoye, 2021), and that may potentially lead to unintended and undesired effects when deployed (Bolukbasi et al., 2016). The bias can stem from the unlabeled texts used for pre-training of Language Models (LMs), or from the language or the label distribution used for tuning a downstream classifier. Since LMs are now the backbone of most NLP model architectures, the extent to which they reflect, amplify, and spread the biases existing in the input data is very important for the further development of such models, and the understanding of their possible harmful outcomes.

Efforts so far have shown a multitude of biases in pre-trained LMs and contextualized embeddings. Sheng et al. (2019) show that pre-training the LM BERT (Devlin et al., 2019) on a medical corpus propagates harmful correlations between genders, ethnicity, and insurance groups. Hutchinson et al. (2020) show that English LMs contain biases against disabilities, where persons with disabilities are correlated with negative sentiment words, and mental illness too frequently co-occur with homelessness and drug addictions. Both Zhao and Bethard (2020) and Basta et al. (2019) show that ELMO (Peters et al., 2018) contains, and even amplifies gender bias. Especially, Basta et al. (2019) discuss the differences of contextualized and non-contextualized embeddings, and which types of gender bias are mitigated and which ones are amplified.

Most work on detecting gender bias has focused on template-based approaches. These templates are simple sentences of the form "`[pronoun] is a [description]`", where a description could be anything from nouns referring to occupations, to adjectives referring to sentiment, emotions, or traits (Stanczak and Augenstein, 2021; Saunders and Byrne, 2020; Bhaskaran and Bhallamudi, 2019; Cho et al., 2019; Prates et al., 2018). Bhardwaj et al. (2021) investigate the propagation of gender biases of BERT in five downstream tasks within emotion and sentiment prediction. They propose an approach to identify gender directions for each BERT layer, and use the Equity Evaluation Corpus (Kiritchenko and Mohammad, 2018) as an evaluation of their approach. They show that their approach can reduce some of the biases in downstream tasks. Nozza et al. (2021) also use a template- and lexicon-based approach, in this case for sentence completion. They introduce a dataset for the six languages English, French, Italian, Portuguese, Romanian, and Spanish, and show that LMs both reproduce and amplify gender-related societal stereotypes.

Another series of work that have focused on template-based datasets are those building on the

| Occupation | Female% | Male% | Occupation | Female% | Male% |
|---|---|---|---|---|---|
| Knitting craftsman | 100 | 0 | Architect | 49.9 | 50.1 |
| Midwife | 99.8 | 0.2 | Lawyer | 48.1 | 51.9 |
| Esthetician | 99.3 | 0.7 | Politician | 48.1 | 51.9 |
| Health Secretary | 98.8 | 1.2 | Associate Professor | 47.2 | 52.8 |
| PhD candidate | 52.8 | 47.2 | Scaffolding builder | 0.5 | 99.5 |
| Psychiatrist | 52.6 | 47.4 | Chief engineer | 0.4 | 99.6 |
| Doctor | 51.6 | 48.4 | Coastal skipper | 0 | 100 |

Table 1: A selection of occupations from the Norwegian statistics bureau, the gold reference distribution of occupations and genders. The occupations presented here are either dominated by more than 98% of either gender, or have a more balanced distribution (underlined percentages) between both female and male genders.

Winograd Schemas data (Levesque et al., 2012). This dataset was developed for the task of coreference resolution, and contains a set of manually annotated templates that requires commonsense reasoning about coreference. It is used to explore the existence of biases in coreference resolution systems, by measuring the dependence of the system on gendered pronouns along stereotypical and non-stereotypical gender associations with occupations. Similarly, the WinoBias (Zhao et al., 2018) dataset focuses on the relationship between gendered pronouns and stereotypical occupations, and is used to explore the existing stereotypes in models. The WinoGender dataset (Rudinger et al., 2018) also contains sentences focusing on the relationship between pronouns, persons, and occupations. Here, they also include gender-neutral pronouns. Unlike WinoBias, WinoGender's sentences are built such that there is a coreference between pronouns and occupations, and between the same pronouns and persons. Based on these datasets for coreference resolution, WinoMT (Stanovsky et al., 2019) has been developed for the task of machine translation. The dataset also contains stereotypical and non-stereotypical templates used to investigate gender bias in machine translation systems.

Moreover, Bender et al. (2021) point out the dangers of LMs and how they can potentially amplify the already existing biases that occur in the data they were trained on. They highlight the importance of understanding the harmful consequences of carelessly using such models in language processing, and how they in particular can hurt minorities. They also discuss the difficulty of identifying such biases, and how complicated it can be to tackle them. This is partly due to poor framework definitions, *i.e.,* how culturally specific they are, but also how unreliable current bias evaluation methods are.

We focus therefore in this work on investigating how culturally specific Norwegian demographics related to gender and occupations are reflected in four Norwegian and two multilingual pre-trained LMs. Our work differs from previous work in that we ground our bias probes to real-world distributions of gender, and rather than expecting the models to always have a balanced representation of genders, we explore to which degree they reflect true demographics.

## 4 Experimental setup

Following the methodology of previous research on gender bias in pre-trained language models, and due to the corresponding lack of resources for Norwegian, we generate our own set of templates that we use with the pre-trained language models to make use of their ability to compute the probabilities of words and sentences. We present an empirical analysis of gender biases towards occupational associations. By using the templates we hope to reduce variation by keeping the semantic structure of the sentence. We analyze the probability distributions of returned pronouns, occupations, and first names; and compare them to real-world gold data representing the demographic distribution in Norway. Investigating the differences between the models can also give us insights into the content of the various types of corpora they were trained on. Data and codes will be made available[1].

Below we discuss in turn (*i*) the gold reference distribution of occupations and genders, (*ii*) the templates, (*iii*) how the templates are used for probing pre-trained language models, and finally (*iv*) the models that we test.

[1] https://github.com/SamiaTouileb/Biases-Norwegian-Multilingual-LMs

**Reference distribution** We use a set of 418 occupations. These represent the demographic distribution of females and males in the respective occupations in Norway[2] originating from the Norwegian statistics bureau. The bureau releases yearly statistics covering various aspects of the Norwegian society, and all data is made freely available. This list comprises a fine-grained level of occupations, where e.g., *lege* (*doctor*) and *allmennlege* (*general practitioner*) are considered two different occupations. The gender-to-occupation ratios in these statistics are used as 'gold standard' when probing the models.

In Table 1 we show some examples of the occupations dominated by more than 98% of either gender, and those that have a more balanced distribution (underlined). Culturally speaking, Norway is known to strive for gender balance in all occupations. While this is true for many instances, there are still some occupations that are unbalanced in gender-distribution. From the Norwegian statistics bureau, it is clear that most midwives are still women, and that most chief engineers are males. However, for occupations as Phd candidates, psychiatrist, doctor, architect, lawyer, politician, and associate professor the distribution of genders is more balanced.

**Templates** Our templates combine occupations, pronouns, and first names. We focus on five template-based tasks, and generate the following corresponding templates that we use as bias probes (Solaiman et al., 2019):

1. Task1: *[pronoun] is a/an [occupation]* (original: *[pronoun] er [occupation]*)
2. Task2: *[pronoun] works as a/an [occupation]* (original: *[pronoun] jobber som [occupation]*)
3. Task3: *[name] is a/an [occupation]* (original: *[name] er [occupation]*)
4. Task4: *[name] works as a/an [occupation]* (original: *[name] jobber som [occupation]*)
5. Task5: *the [occupation] [name]* (original: *[occupation] [name]*)

As pronouns, our work mainly focuses on *hun* and *han* (*she* and *he* respectively). As demographic statistics are still made using a binary gender distribution, we could not include the gender neutral pronoun *hen* (*they*), which is, in addition, rarely used in Norway.

As first names, we also extract from the Norwegian statistics bureau[3] the 10 most frequent female and male names in Norway from 1880 to 2021, this results in 90 female names and 71 male names. For tasks 1–4 we use the full set of 418 occupations, while in task 5 we focus on those that either have a balanced distribution between genders or are clearly female- or male-dominated. This was decided after an analysis of the distribution of occupations across genders, and resulted in two thresholds. All occupations that had between 0 and 10% differences in distribution, were deemed balanced (*e.g.,* 51% female and 49% male). All occupations that had more than 75% distribution of one gender against the other, were deemed unbalanced, and are referred to as either clearly female ($\geq$75%) or clearly male ($\geq$75%) occupations. This resulted in a set of 31 clearly female occupations, 106 clearly male occupations, and 49 balanced occupations.

For tasks 1 and 2, we mask the pronouns and compute the probability distribution across the occupations for female and male pronouns. For tasks 3, 4, and 5, we mask the occupations and compute the probability distributions in each bias-probe. Masking pronouns will allow us to uncover how likely a gendered pronoun is correlated with an occupation, and masking the occupation will allow us to uncover how likely occupations are correlated with female and male names.

**Probing and evaluation** For each task, we first generate the probability distributions of masked tokens in each bias probe. In order to have a comparable distribution to the gold standard (which is given as a percentage), we compute a simple percentage representation of the probability distributions by following the following formula:

$$\text{f\_pron\%} = \frac{\text{prob f\_pron}}{\text{prob f\_pron+prob m\_pron}}$$

Where *f_pron%* is the percentage of a female pronoun, and *prob x_pron* is the output probability of each model for each of the female and male pronouns. The same simple formula is used in all tasks. We are aware that this is a simplified representation of the output of each model, nevertheless, we believe that it will not change the overall distribution.

Once probability distributions are mapped to per-

[2] https://utdanning.no/likestilling

[3] https://www.ssb.no/befolkning/navn/statistikk/navn

centages, we quantify the difference between female and male scores by simply subtracting the scores of males from the scores of female. Positive values will represent occupations that are more strongly associated with females than males by the model, and negative values represent the opposite. This is also applied to the gold standard data. We use the demographic distribution of the occupations from the Norwegian statistics bureau as gold data.

Based on this, values greater than 0 are deemed female-dominated occupations, and values lower that 0 are male-dominated occupation. This is used to compute the macro F1 values for each model.

**Pre-trained language models** We analyse the predictions of six pre-trained language models, four Norwegian and two multilingual. Note that Norwegian has two official written standards; Bokmål (literally 'book tongue') and Nynorsk (literally 'new Norwegian'). While Bokmål is the main variety, roughly 15% of the Norwegian population write in the Nynorsk variant. All the Norwegian models are trained on data comprising both Bokmål and Nynorsk.

- NorBERT (Kutuzov et al., 2021): trained on the Norwegian newspaper corpus[4], and Norwegian Wikipedia, comprising about two billion word tokens.
- NorBERT2[5]: trained on the non-copyrighted subset of the Norwegian Colossal Corpus (NCC)[6] and the Norwegian subset of the C4 web-crawled corpus (Xue et al., 2021). In total, it comprises about 15 billion word tokens.
- NB-BERT (Kummervold et al., 2021): trained on the full NCC, and follows the architecture of the BERT cased multilingual model (Devlin et al., 2019). It comprises around 18.5 billion word tokens.
- NB-BERT_Large[7]: trained on NCC, and follows the architecture of the BERT-large uncased model.
- mBERT (Devlin et al., 2019): pre-trained on a set of the 104 languages with the largest Wikipedia pages.
- XLM-RoBERTa (Conneau et al., 2020): trained on a collection of 100 languages from the Common Crawl corpus.

As can be seen above, each model has been trained on different types of corpora, and are all of various sizes. The NCC corpus, is a collection of OCR-scanned documents from the Norwegian library's collection of newspapers and works of fiction (with publishing years ranging from early 1800s to present day), government reports, parliament collections, OCR public reports, legal resources such as laws, as well as Norwegian Wikipedia. In short, some models are trained on well structured texts, that follow a somewhat formal style, while other models also include less structured texts in the form of online content.

## 5 Results

Table 2 summarizes the overall results for all models. We also compute class-level F1 values for each task, these can be found in Table 3 and Figure 5. Below we discuss the task-wise results in more detail.

### 5.1 Task1: (she|he) is a/an [occupation]

In the first task, we mask the pronouns *she* and *he* in our bias probes. We focus on the full set of 418 occupations. As can be seen in Table 2, all four Norwegian models give higher scores than the two multilingual models. NB-BERT and NB-BERT_Large have a macro F1 of 0.75, and are the highest performing models overall. It should be pointed out that these are also the biggest Norwegian models in terms of token counts. NorBERT is the less performing Norwegian model in this task, and has a macro F1 a few percentiles higher than the multilingual model XLM-RoBERTa. We believe that this might be impacted by the the size of NorBERT, which is the smallest Norwegian model in terms of token counts.

Looking at class-level F1 scores from Table 3, all models achieve high F1 scores for the male class, with NB-BERT_Large achieving the highest score with an F1 of 0.84, and mBERT achieving the lowest one with an F1 of 0.74. In contrast, all models have substantially lower F1 score on the female class. Again, NB-BERT_Large achieves the highest score with 0.67 F1, and mBERT the lowest with 0.30. This shows that the models are already somehow skewed towards the male class.

[4] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/

[5] https://huggingface.co/ltgoslo/norbert2

[6] https://github.com/NbAiLab/notram/blob/master/guides/corpus_description.md

[7] https://huggingface.co/NbAiLab/nb-bert-large

| model | Task1 | Task2 | Task3 | Task4 | Task5_b | Task5_ub |
|---|---|---|---|---|---|---|
| NorBERT | 0.69 | 0.67 | 0.60 | 0.35 | 0.46 | **0.83** |
| NorBERT2 | 0.73 | 0.54 | 0.77 | 0.72 | 0.52 | 0.76 |
| NB-BERT | **0.75** | 0.74 | 0.70 | **0.80** | **0.69** | 0.77 |
| NB-BERT_Large | **0.75** | **0.82** | **0.80** | 0.74 | 0.49 | 0.76 |
| mBERT | 0.52 | 0.42 | 0.52 | 0.52 | 0.52 | 0.55 |
| XLM-RoBERTa | 0.65 | 0.50 | 0.68 | 0.49 | 0.47 | 0.56 |

Table 2: Macro F1 of models compared to the real-world "gold" distribution. **Task1**: `[pronoun] is a/an [occupation]`, **Task2**: `[pronoun] works as a/an [occupation]`, **Task3**: `[name] is a/an [occupation]`, **Task4**: `[name] works as a/an [occupation]`, **Task5_b**: `the [occupation] [name]` with balanced distributions in gold, **Task5_ub**: `the [occupation] [name]` with clearly female and male occupation distributions in gold.

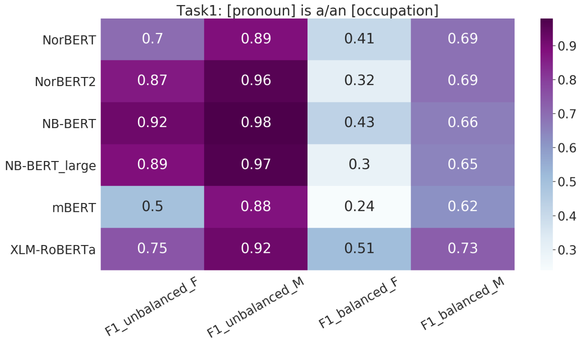

Figure 1: Task1, class-level F1 values focusing on balanced and unbalanced occupations.

In addition to looking at the distribution of all occupations, and based on the previous observation that all models seem to reflect male occupations but to a lesser extent reflect female occupations, we have looked at the occupations that have balanced and unbalanced distributions in the gold data. The unbalanced occupations as previously mentioned, are those which are clearly female or male occupations (more than 75% distribution of one gender against the other). The balanced distribution are those that have between 0 and 10% differences in gender distribution in the gold data. Results are depicted in Figure 1.

When it comes to clearly female occupations, the three biggest Norwegian models, namely NorBERT2, NB-BERT, and NB-BERT_Large obtain highest F1 values with 0.87, 0.92, and 0.89 respectively. Followed by XLM-RoBERTa and NorBERT. For clearly male occupations, all models have high F1 values, with the three top ones being again NorBERT2, NB-BERT, and NB-BERT_Large. The two multilingual models achieve quite high values, with XLM-RoBERTa outperforming NorBERT here again. It is quite clear that the Norwegian models have a good representation of clearly female and male occupations. Another compelling result is that XLM-RoBERTa has a quite accurate representation of these unbalanced occupations, equating the ones from the smallest Norwegian model NorBERT.

Focusing on balanced occupations, most models exhibit a tendency to represent occupations as male. NorBERT, NB-BERT, and XLM-RoBERTa are the only models that seem to have a decent representation of female occupations. The expectations here are not that the models would give a better representation of female occupations, but rather be equally good at representing both genders.

## 5.2 Task2: (she|he) works as a/an [occupation]

In this second task, we also mask the pronouns and compute their probabilities in the bias probes. We here again focus on the full set of occupations, 418 occupations.

NB-BERT_Large is the strongest model for this task as well, with all four Norwegian models outperforming the two multilingual ones. Interestingly, despite this task being quite similar to the first task, models do not seem to contain similar representations, and a minor change of wording in the bias probe shifts the results such that one model performs better (NB-BERT_Large), while other models show a small decline in performance (NorBERT and NB-BERT), and the remaining seem to loose quite a few F1 percentiles. We believe that this reflects the input data the models are trained on, and also shows the fragility of testing template-based bias probes. Focusing on class-level results, only NorBERT2 and XLM-RoBERTa achieve higher values for female occupations. The rest of the mod-

|  | Task1 |  | Task2 |  | Task3 |  | Task4 |  |
|---|---|---|---|---|---|---|---|---|
| model | F | M | F | M | F | M | F | M |
| NorBERT | 0.59 | 0.78 | 0.57 | 0.77 | 0.61 | 0.60 | 0.58 | 0.13 |
| NorBERT2 | 0.63 | 0.83 | 0.63 | 0.45 | 0.71 | **0.84** | 0.72 | 0.71 |
| NB-BERT | 0.66 | 0.83 | 0.73 | 0.74 | 0.60 | 0.81 | **0.77** | **0.84** |
| NB-BERT_large | **0.67** | **0.84** | **0.77** | **0.87** | **0.77** | 0.82 | 0.74 | 0.74 |
| mBERT | 0.30 | 0.74 | 0.07 | 0.76 | 0.34 | 0.69 | 0.31 | 0.73 |
| XLM-RoBERTa | 0.52 | 0.77 | 0.60 | 0.40 | 0.59 | 0.76 | 0.61 | 0.36 |

Table 3: Class-level (Male/Female) F1 when compared to the real-world "gold" distribution for tasks 1–4

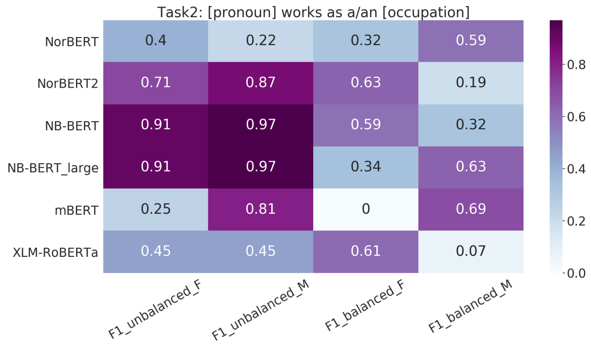

Figure 2: Task2, class-level F1 values focusing on balanced and unbalanced occupations.

els mostly represent male occupations, except for NB-BERT, which seems to be equally good at representing both.

Similarly to Task1, we did a more thorough analysis by focusing on the balanced and unbalanced distributions of occupations, this can be seen in Figure 2.

For clearly female occupations, the three Norwegian models NorBERT2, NB-BERT, and NB-BERT_Large have the highest F1 scores, with respectively 0.71, 0.91, and 0.91. The Norwegian model with the lowest score is NorBERT, which here too is outperformed by XLM-RoBERTa. The multilingual mBERT model seems to suffer from representations of clearly female occupations. Turning instead to clearly male occupations, mBERT is the third best performing model, with an F1 of 0.81, preceded by NorBERT2 with 0.87 F1, and NB-BERT and NB-BERT_Large with both an F1 of 0.97. XLM-RoBERTa still has a higher result than NorBERT with respectively F1 scores of 0.45 and 0.22. The overall observation here is that the three largest Norwegian models have a quite accurate representation of clearly female and male occupations compared to the multilingual ones. It also seems that the size of the training data matters, as NorBERT does not equate with other models.

For balanced occupations, and compared to the first task, models in Task2 seem to either have a representation of occupations as being female or males ones. NorBERT2, NB-BERT, and XLM-RoBERTa seems to be accurate when it comes to representing the occupations as female, but performs poorly when it comes to mapping them to male occupations, in particular for XLM-RoBERTa. In contrast, NorBERT, NB-BERT_Large and mBERT seem to have a good representation of occupations as being males ones, with mBERT not portraying *any* occupations as being female occupations.

### 5.3 Task3: [name] is a/an [occupation]

In this task, we use the set of most frequent Norwegian first names from 1880 to 2021. Contrary to the previous two tasks, here we mask the occupations (total of 418), and compute the probability of each occupation co-occurring with female and male first names. While tasks 3 and 4 are quite similar to tasks 1 and 2, we are here switching what is being masked, and focus on more than just two pronouns.

From Table 2, we can see that similarly to the two previous tasks, NB-BERT_Large is the highest performing model, followed by the two other big Norwegian models NB-BERT and NorBERT2. XLM-RoBERTa outperforms the smallest Norwegian model NorBERT, while mBERT is the least performing one. The results for this task are comparable to the most similar task, Task1.

Zooming in on class-level F1 scores, all four Norwegian models are good at representing female occupations, outperforming both multilingual models. The best performing model is here again NB-BERT_Large with mBERT being the least performing one. For male occupations, all models achieve high scores, with NorBERT2 achieving the high-

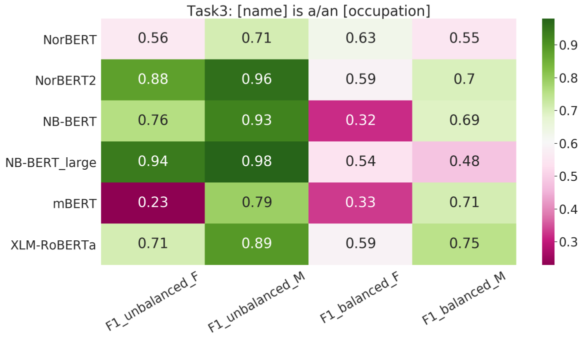

Figure 3: Task3, class-level F1 values focusing on balanced and unbalanced occupations.

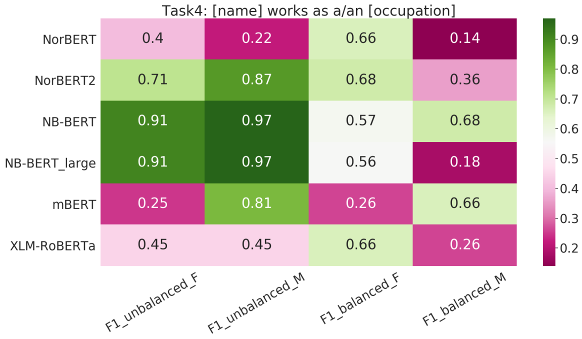

Figure 4: Task4, class-level F1 values focusing on balanced and unbalanced occupations.

est F1 of 0.84, and NorBERT achieving the lowest score of 0.60 F1.

As for the two previous tasks, we also look at the balanced and unbalanced occupations from the gold data, and explore how each of these are reflected in the models using Task3's bias probe. These can be seen in Figure 3.

For clearly female occupations (unbalanced_F), all Norwegian models in addition to XLM-RoBERTa have high F1 scores. Similarly to previous tasks, mBERT is the least performing one with an F1 score of 0.23. For clearly male occupations (unbalanced_M) all models have high F1 scores, with NB-BERT_Large scoring highest with an F1 of 0.98, followed by NorBERT2 (0.96), NB-BERT (0.93), XLM-RoBERTa (0.89), mBERT (0.79), and NorBERT (0.71). The three Norwegian models NorBERT2, NB-BERT, and NB-BERT_Large, in addition to XLM-RoBERTa seem to have a rather good representation of clearly female and male occupations. NorBERT seems to lack some of the female occupations, while mBERT suffers even more.

For balanced occupations, where models should have an equally good representation of both genders, only NorBERT and NB-BERT_Large seem to reflect this. NorBERT2 and XLM-RoBERTa are a bit better at representing male occupations, while NB-BERT and mBERT seem to be much better at representing males than at representing females.

### 5.4 Task4: [name] works as a/an [occupation]

Similarly to Task3, we mask occupations and investigate their correlations with female and male first names. As for Task2, we here use the probe fixed by the sequence "works as a/an". From Table 2, it is apparent that the three big Norwegian models NorBERT2, NB-BERT, and NB-BERT_Large with respective F1 scores of 0.72, 0.80, 0.74, are the models with the highest scores for the task. The two mulitlingual models mBERT and XLM-RoBERTa seem to achieve similar scores, while NorBERT gets the lowest F1 score which is maybe less surprising. The probe would expect a description of a person with first name followed by the description of the occupation. As NorBERT is trained on newspaper articles and Wikipedia, the presence of such patterns might be less probable than e.g. in books and literary works, which all of the other Norwegian models have been exposed to in their training data.

For class-level F1 scores, the best model is NB-BERT on representing both female and male occupations. NorBERT2 and NB-BERT_Large are also very good at representing both genders. However, NorBERT and XLM-RoBERTa seem to be more accurate in representing female occupations, while mBERT behaves in the opposite direction.

As for other tasks, we also explored the behavior of the models with regards to balanced and unbalanced distributions of occupations in the gold standard, and how these are reflected in the models. This can be seen in Figure 4.

Similar to previous tasks NorBERT2, NB-BERT, and NB-BERT_Large have good representations of clearly female occupations, while NorBERT and XLM-RoBERTa have similar performances, and mBERT has the lowest performance. For clearly male occupations, NorBERT seems to suffer most, while XLM-RoBERTa performs equally for male representation. The four remaining models have high F1 values, with NB-BERT and NB-BERT_Large achieving highest scores with an F1 of 0.97. For balanced occupations, NorBERT, NorBERT2, NB-BERT_Large, and XLM-RoBERTa have decent F1 scores and seem to represent occupations as female ones. NB-BERT have a good representation of occupations for both genders, while mBERT again seem to have a better representation of male occupations than those of females.

### 5.5 Task5: the [occupation] [name]

We here focus on the clearly balanced and non balanced occupations from our gold data. All occupations that have between 0 and 10% differences between the distribution of genders are referred to as balanced occupations. Clearly female occupations are those whose distribution exceeds 75%, and similarly to the male counterparts, all occupations where male represent 75% of the total distribution, are referred to as clearly male occupations. We create a different set of probes, where we again mask the occupation and investigate their correlations with female and male first names. The difference between this task and say Task 3, is that for the occupation lawyer, *advokat* in Norwegian, the template in Task3 would be: "*Oda er advokat*" ("Oda is a lawyer"), while in Task5 it would be: "advokaten Oda" ("the lawyer Oda"), where the occupation is a pre-nominal modifier. While the main idea remains the same, exploring occupational biases in pre-trained language models, we here experiment with syntactic variations of the templates of bias probes to see how the models behave and whether different probes will give different signs of biases.

Focusing on the balanced occupations, from Table 2, all models achieve an F1 score of at least 0.46, with NB-BERT reaching the highest F1 value of 0.69. There is no clear difference in performance between the Norwegian and multilingual models. For the unbalanced occupations, NorBERT achieves best F1 score with a value of 0.83. Followed by NB-BERT, NorBERT2, and NB-BERT_Large with respectively 0.77, 0.76, and 0.76 F1 values. While the two multilingual models have at least 0.20 F1 values less than the least performing Norwegian model. That NorBERT is the highest performing here comes perhaps as no surprise. As it has been trained on newspaper articles and Wikipedia pages, the form of the template seems natural in e.g. reporting cases where people are introduced by their occupations.

Class-based F1 scores can be seen in Figure 5. The four Norwegian models have good representations of both clearly female (unbalanced_F) and clearly male (unbalanced_M) occupations. With NorBERT achieving higher scores on both genders, and being the best model. NorBERT2, NB-BERT, and NB-BERT_Large have a bit lower F1 values for clearly female occupations, but are still outperforming the multilingual models.

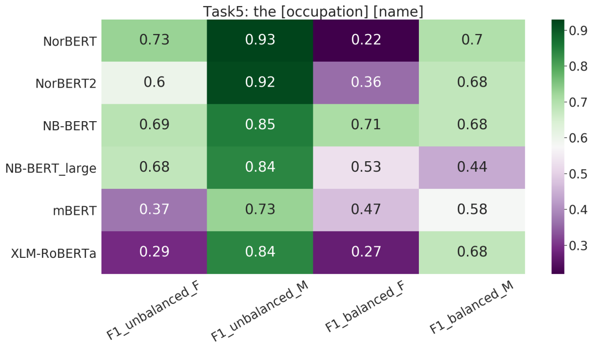

Figure 5: Task5, class-level F1 values focusing on balanced and unbalanced occupations.

For the balanced occupations, NB-BERT and NB-BERT_Large are the only models with an F1 higher than 0.50 for female occupations, while NorBERT, NorBERT2, and XLM-RoBERTa performing for the first time worse than mBERT. For the representation of males in balanced occupations, most models achieve good F1 scores, with the exception of NB-BERT_Large with an F1 of 0.44. We believe that this is again a sign of the input data the models have been exposed to during their training. Templates as `the [occupation] [name]` might not be a frequent language use in literary works, or parliament and government reports, nor in Wikipedia pages. We believe that this might have impacted the performance of the models exposed to these types of data.

## 6 Discussion

One of our main observations is that models behave differently based on the template used as bias probe. The templates we have used, in *e.g.,* Task1 and Task2, and Task3 and Task4, differ only by one token, and do not change the semantics of the template even if it changes its syntactic realization. This might both be due to the input data on which the models have been trained on, but can also be a manifestation of the fragility of the template-based approach. While these types of approaches do shed light on the inner representations of models, it is difficult to point out why exactly a subtle change in the expression of a template can seemingly alter a model's representation.

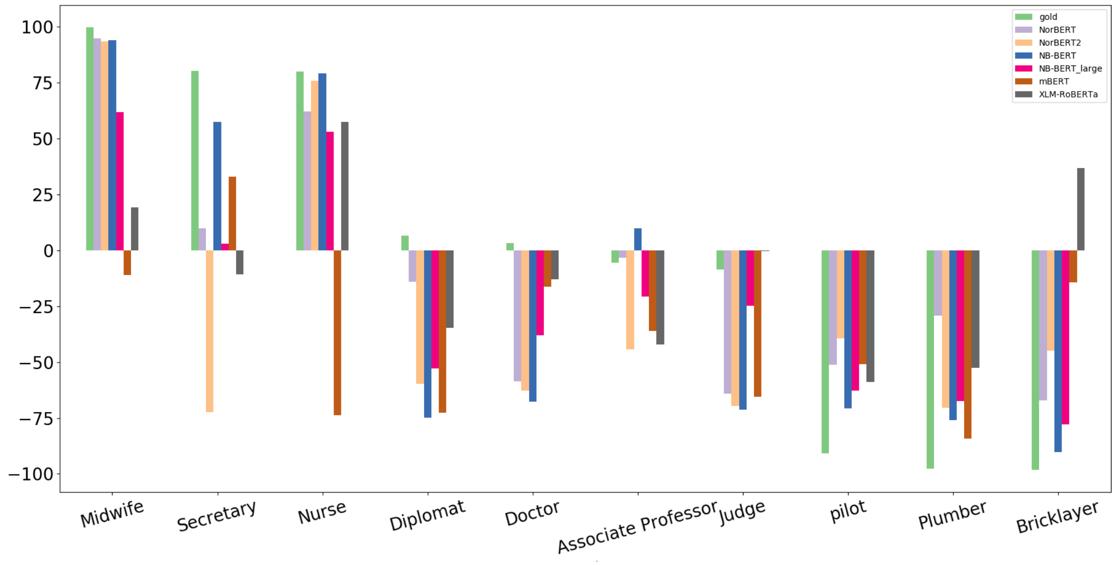

Figure 6: Example of balanced and unbalanced occupations in gold data, and each model's prediction in Task1.

Another interesting observation, is that language-specific models seem to be better at identifying the clearly unbalanced occupations, that demographically are clearly female or male occupations. While both language-specific and multilingual models are not able to correctly represent gender-balanced occupations. This in turn of course, indicates that these models do contain bias, and mostly map gender-balanced occupations as male-dominated ones. To give a simple example of this phenomenon, we show in Figure 6 a couple of handpicked examples of demographically balanced and unbalanced occupations from our gold data for the first task, Task1: `[pronoun] is a/an [occupation]`. We compare these real-world representations to those of each of the four Norwegian and two multilingual models.

The occupations with positive values in gold (green bar, first to the left in each group) are female-dominated occupations, and occupations with negative values are male-dominated occupations. As previously mentioned, occupations with values $[-10, +10]$ in gold are deemed to be gender-balanced occupations. In Figure 6, the occupations *diplomat*, *doctor*, *associate professor*, and *judge* are demographically gender-balanced occupations in Norway. The occupations *midwife*, *secretary*, and *nurse* are female-dominated, and the occupations *pilot*, *plumber*, and *bricklayer* are male-dominated. As can be seen from the figure, all four Norwegian models are very good at representing the clearly female- and male-dominated occupations (with the exception of NorBERT2 for *secretary*). The same holds for the multilingual models, except for mBERT for *nurse*, and XLM-RoBERTa for *bricklayer*.

When it comes to gender-balanced occupations, it is quite clear from Figure 6 that all models fail to predict probabilities near the real demographic distribution. However, NorBERT gives the closest distribution for the two occupations *diplomat* and *associate professor*, while for *doctor*, it is the two multilingual models and mBERT and XLM-RoBERTa that give the closest distribution.

## 7 Conclusion

We have presented in this paper an investigation into how a demographic distribution of occupations, along two gender dimensions, is reflected in pre-trained language models. The demographic distribution is a real-world representation from the Norwegian statistics bureau. Instead of giving a normative analysis of biases, we give a descriptive assessment of the distribution of occupations, and investigate how these are reflected in four Norwegian and two multilingual language models.

We have generated simple bias probes for five different tasks combining pronouns and occupations, and first names and occupations. Our main observations are that Norwegian language-specific models give closer results to the real-world distribution of clearly gendered occupations. Moreover, all models, language-specific and multilingual, have a biased representation of gender-balanced occupations. Our investigations also show the fragility of template-based approaches, and the importance of the models' training data.

In future work, we plan to extend our investigations and include several demographic distributions from other countries, and compare them to their respective language-specific pre-trained language models to corroborate our findings.

## Acknowledgment

This work was supported by industry partners and the Research Council of Norway with funding to *MediaFutures: Research Centre for Responsible Media Technology and Innovation*, through the centers for Research-based Innovation scheme, project number 309339.

## References

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 *Proc. of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation*, 13(4).

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy. Association for Computational Linguistics.

Su Lin Blodgett. 2021. Sociolinguistically driven approaches for just natural language processing.

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.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8).

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Per Egil Kummervold, Javier de la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a norwegian transformer model. In *Proc. of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for norwegian. In *Proc. of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2018. Assessing gender bias in machine translation – a case study with google translate.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Addressing exposure bias with document minimum risk training: Cambridge at the WMT20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 862–869, Online. Association for Computational Linguistics.

Krunal Shah, Nitish Gupta, and Dan Roth. 2020. What do we expect from multiple-choice QA systems? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3547–3553, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

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*, 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.

Yiyun Zhao and Steven Bethard. 2020. How does BERT's attention change when you fine-tune? an analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics.