Exploring the Effects of Negation and Grammatical Tense on Bias Probes

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

Abstract

We investigate in this paper how correlations between occupations and gendered-pronouns can be affected and changed by adding negation in bias probes, or changing the grammatical tense of the verbs in the probes. We use a set of simple bias probes in Norwegian and English, and perform 16 different probing analysis, using four Norwegian and four English pre-trained language models. We show that adding negation to probes does not have a considerable effect on the correlations between gendered-pronouns and occupations, supporting other works on negation in language models. We also show that altering the grammatical tense of verbs in bias probes do have some interesting effects on models' behaviours and correlations. We argue that we should take grammatical tense into account when choosing bias probes, and aggregating results across tenses might be a better representation of the existing correlations.

1 Introduction

Pre-trained Language Models (LMs) reflect various linguistic and factual knowledge, represented in the data they have been trained or fine-tuned on. Despite their emergent success, these LMs might contain various degrees of representational harms, where genders, religions, and ethnicity might be miss-represented, or not represented at all (Blodgett et al., 2020; Bender et al., 2021).

LMs can contain biases that might be inherited by the unlabeled data used while training them, the data used while fine-tuning them, and the label distribution used for downstream classifiers. In recent years, the extent to which these LMs reflect, amplify, and spread the biases existing in the input data has been an active research focus as it is important to understand their inner representations, and what can be their possible harmful outcomes. The possible harmful effects of LMs have been thoroughly discussed by Bender et al. (2021), especially their ability to potentially amplify the already existing biases that occur in the data they were trained on.

Some of the efforts so far have demonstrated the existence of different types of biases that correlate gender and ethnicity with insurance groups (Sheng et al., 2019), people with disabilities and mental illnesses with negative sentiment words, homelessness, and drug addictions (Hutchinson et al., 2020), and that they can even amplify gender bias (Zhao and Bethard, 2020; Basta et al., 2019)

One way to explore the existence, and types, of gender bias in LMs is to use template-based approaches (Stanczak and Augenstein, 2021; Saunders and Byrne, 2020; Bhaskaran and Bhallamudi, 2019; Cho et al., 2019; Prates et al., 2018). These template-based approaches have for example been used to show how LMs can reproduce and amplify gender-related societal stereotypes (Nozza et al., 2021), and how the gender biases in BERT propagate in tasks within emotion and sentiment prediction (Bhardwaj et al., 2021).

Moreover, these LMs when queried using template-based probes, seem to not distinguish between templates and their negation (Kassner and Schütze, 2020), and therefore suggesting that they are not always able to handle negation. Kassner and Schütze (2020) have also explored perturbing the probes by adding misprimes to extract information from LM, and showed that LMs are sensitive. The fragility of the template-based probes has also been pointed out by Touileb et al. (2022), where they have shown that sometimes a simple word change can alter a model's behaviour.

In this paper, we investigate the effects of negation and grammatical tense when probing LMs for gender bias purposes. Based on previous investigations, and research on probing language models, our main hypothesis is that changing the formulation of a probe can have an effect on the output of a LM. We know that LMs use datasets of vari-

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 423–429
November 20–23, 2022. ©2022 Association for Computational Linguistics

	Norwegian	English
present	*[pronoun] jobber som [occupation]*	*[pronoun] works as a/an [occupation]*
past	*[pronoun] jobbet som [occupation]*	*[pronoun] worked as a/an [occupation]*
future	*[pronoun] skal jobbe som [occupation]*	*[pronoun] will work as a/an [occupation]*
future	*[pronoun] kommer til å jobbe som [occupation]*	*[pronoun] is going to work as a/an [occupation]*
N. present	*[pronoun] jobber ikke som [occupation]*	*[pronoun] does not work as a/an [occupation]*
N. past	*[pronoun] jobbet ikke som [occupation]*	*[pronoun] did not work as a/an [occupation]*
N. future	*[pronoun] skal ikke jobbe som [occupation]*	*[pronoun] will not work as a/an [occupation]*
N. future	*[pronoun] kommer ikke til å jobbe som [occupation]*	*[pronoun] is not going to work as a/an [occupation]*

Table 1: Bias probes altered with grammatical tense and negation. "N." stands for "negated". We focus on binary gendered-pronouns, and use a set of occupations from the Norwegian statistics bureau.

ous sizes, that cover various time-periods, and that these time-periods can reflect different perspectives on society and how genders can be correlated with occupations. Using probes in past tense might only reflect how a gender used to be correlated with some occupations, discarding other correlations that might be expressed using future tense. The same for negation, even if empirical evidence have shown that it is not well handled by LMs (Kassner and Schütze, 2020).

We explore four Norwegian and four English LMs using simple probes related to occupations, in correlation with pronouns. First, we alter the probes by adding negation and comparing the scores attributed to the pronouns. We thereafter alter the grammatical tense of the verb in our probes, and again compare the scores of the pronouns attributed by each model. More precisely, we focus on exploring the following questions:

- What is the effect of negating or changing the grammatical tense of a bias probe?
- What effect do these changes have on the correlations of gendered-pronouns with occupations?

To address these questions, we inspect how sensitive bias probes can be, and analyse the effects of our experiments on the behaviours of Norwegian and English pre-trained LMs. We start in Section 2 by describing our experimental setup, give details about our bias probes, and the LMs used. In Section 3 we present and discuss our main results and findings. Finally, in Section 4, we conclude and summarize our work, and discuss some possible future work.

2 Experiments

We use the definition of bias by Friedman and Nissenbaum (1996), where bias is the systematic discrimination against, and unfairly process of, a certain group of individuals exhibited by automated systems. In this work, we look at the correlations within the pre-trained models between gendered pronouns and professional occupations, and explore how the scores returned by the LMs can change by simple alterations in the probes. In our case, introducing negation and altering the grammatical tense of the verbs. However, we do not evaluate if a model is biased or not, we rather look at what changes when the probes are perturbed. We do not try to reduce the stereotypical representations, but rather shed light on how fragile, sensitive, or reliable the bias probes are.

We use the masked-language modeling objective of each model to predict the probability of pronouns in a probe. For simplicity, we also do not look at the degree of variation in the returned probabilities, but we simply check which pronoun has a greater value, and use this prediction to analyse the effect of the negated and tense-specific probes.

One limitation of our work is that we only look at the correlations between occupations and binary gender categories (male and female), although we acknowledge the fact that gender as an identity spans a wider spectrum.

2.1 Bias probes

The templates we use combine a set of occupations with gendered pronouns. The occupations we use are from the Norwegian statistics bureau[1], and are at a fine-grained level, such that *lege* (*doctor*) and *allmennlege* (*general practitioner*) are considered two different occupations. We select the set of 353 occupations that we define as statistically *clearly* female or male occupations. These are the occupations that have a statistical difference of more than 15% between genders. We also translate these

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

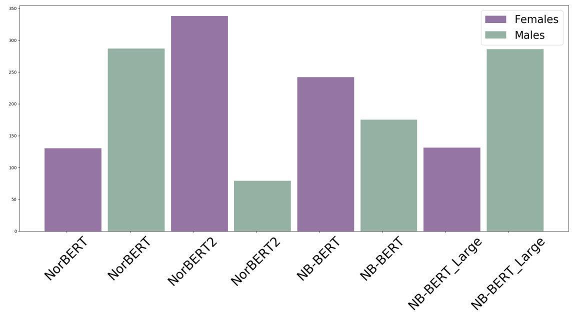

Figure 1: Correlations of genders with occupations for the bias probe "*[pronoun] jobber som [occupation]*" in Norwegian language models.

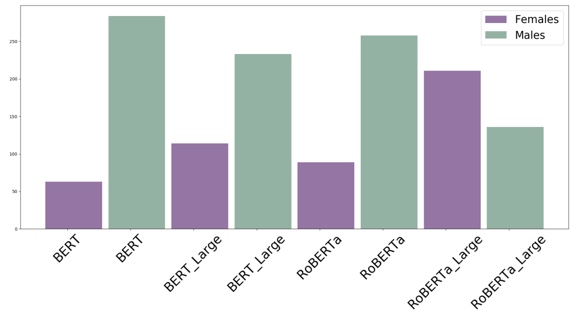

Figure 2: Correlations of genders with occupations for the bias probe "*[pronoun] works as [occupation]*" in English language models.

occupations to English, in order to use them with the English models. Both the list of Norwegian and English occupations are made available.[2]

We base our work on two probes, one in Norwegian (*[pronoun] jobber som [occupation]*) and it's equivalent in English (*[pronoun] works as a/an [occupation]*). Based on these two, we generate three additional probes per language representing past and future forms, resulting in four probes per language. We then generate the negated versions of these probes, resulting in eight probes in total. The full list of probes can be seen in Table 1.

When it comes to pronouns, and as previously mentioned, we focus on a binary representation using the English pronouns "she" and "he" and their Norwegian equivalent "hun" and "han", .

2.2 Models

We inspect the predictions of eight pre-trained language models, four for each language.

Norwegian models Norwegian has two official written standards: Bokmål and Nynorsk. All the Norwegian models are trained on data comprising both written standards. The models we use are:

- NorBERT (Kutuzov et al., 2021): trained on the Norwegian newspaper corpus[3], and Norwegian Wikipedia.
- NorBERT2[4]: trained on the non-copyrighted subset of the Norwegian Colossal Corpus (NCC)[5] and the Norwegian subset of the C4 web-crawled corpus (Xue et al., 2021).
- NB-BERT (Kummervold et al., 2021): trained on the full NCC. Distinctively from the two previous models, follows the architecture of the multilingual BERT cased model (Devlin et al., 2019).
- NB-BERT_Large[6]: trained on NCC, and based on the architecture of the BERT-large uncased model.

English models For the English models we use BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), both in their *base* and *large* forms. We chose to focus on these models, instead of more recent English models, because their architectures are more similar to the Norwegian ones. Both models have also been shown to contain various types of biases (Sheng et al., 2019).

3 Results and Discussion

The two original probes in present, non-negated, forms are "*[pronoun] jobber som [occupation]*" for Norwegian, and "*[pronoun] works as a/an [occupation]*" for English. In Figures 1 and 2 we show the distribution of gendered-pronouns based on the returned probabilities of the Norwegian and English LMs. The y axis here is the number of occupations correlated with each gendered-pronoun, in each model, when using the bias probes.

As can be seen in Figure 1, the models NorBERT and NB-BERT_Large tend to heavily correlate occupations with male gender. While it seems to be the opposite for NorBERT2 and NB-BERT. This however does not hold for the English models. Ex-

[2] https://github.com/SamiaTouileb/Sensitivity-of-Bias-Probes

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

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

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

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

comparison	Total shift	shifted to F	Shifted to M	Total shift	shifted to F	Shifted to M
	NorBERT			NorBERT2		
present VS past	20.39%	0%	100%	33.71%	2.52%	97.47%
present VS future	16.99%	98.33%	1.66%	12.18%	34.88%	65.11%
present VS future2	9.63%	58.82%	41.17%	15.29%	12.96%	87.03%
	NB-BERT			NB-BERT_Large		
present VS past	9.91%	2.85%	97.14%	5.66%	85%	15%
present VS future	14.44%	94.11%	5.88%	7.08%	68%	32%
present VS future2	16.14%	100%	0%	7.08%	80%	20%
	BERT			BERT_Large		
present VS past	8.35%	0%	100%	17.00%	0%	100%
present VS future	2.88%	80%	20%	7.49%	42.30%	57.69%
present VS future2	4.03%	14.28%	85.71 %	6.05%	42.85%	57.14%
	RoBERTa			RoBERTa_Large		
present VS past	9.51%	6.06%	93.93%	7.20%	8%	92%
present VS future	10.08%	5.71%	94.28%	8.93%	19.35%	80.64%
present VS future2	10.95%	10.52%	89.47%	10.37%	41.66%	58.33%

Table 2: Percentage of occupations that have shifted correlations from one gender to another, by changing the verb tense in the bias probes. Such that: present (*jobber som|works as a/an*), past (*jobbet som|worked as a/an*), future (*skal jobbe som|will work as a/an*), and future2 (*kommer til å jobbe som|is going to work as a/an*).

cept for RoBERTa_Large, all the other three models correlate most occupations with male gender.

It is based on these distributions that we build our analysis. We do not analyse which occupations are correlated with male and females, we rather quantify how many females and males are represented in each probe, and how that changes when perturbing the probes.

It has already been shown that LMs do not handle negation that well (Kassner and Schütze, 2020). Our analysis of bias probes and how they behave with regards to negation also supports this claim. By looking at the distribution of female and male correlated occupations using our eight negated bias probes, it is apparent that all models, return somewhat the same correlations between occupations and genders. Very few models exhibit changes in the correlations: 24 out of 32 combinations of probes and models show a shift in less than 16% of occupations. This shows that negation have little effect on bias probes, and rarely changes the correlations between genders and occupations. See Tables 3 and 4 in Appendix A for the statistical distributions of these results.

Some interesting observations can also be made when it comes to altering the grammatical tense of probes. Table 2 shows the percentage of the total number of occupations that have shifted correlations from one gender to another, for each Norwegian and English LMs, and for all our bias probes. We also give a breakdown of percentages into occupations that have shifted correlations to either gender.

Interestingly, shifting the tense from present to past tense seems to shift the correlations between occupations and genders towards male pronouns. This observation holds for all English and Norwegian models, but does not apply for the biggest Norwegian model NB-BERT_Large.

When shifting the tense from present to future, the opposite seems to happen. The changes seem to mainly shift the correlations of occupations from males to females. This is true for most Norwegian models (except NorBERT2), but does not hold for the English models (except for BERT – see Table 2). These changes in correlations are a sign of the sensitivity of the template-based probe approach. Altering the probes can change the models' behaviours, and in a simple analysis like this, change the overall distribution of correlations between genders and occupations.

The same observations can be seen with the negated tense probes. All Norwegian models shift correlations to male-gendered pronouns when switching from present to past tense, while shifting to female-gendered pronouns if comparing probes between present and future tense. For the English models, all seem to change the correlations towards male-gendered pronouns when shifting tenses except for two instances of "present VS future" for the models BERT_Large and RoBERTa. For more details about this, see Table 5 in Appendix A.

We think that one possibility for the differences between the observations made on the Norwegian and English models is the name of the occupations. As these were selected from the Norwegian statistics bureau, they might reflect Norwegian demographics more than the English models. Some of the fine-grained occupations might not be as frequent in English-speaking countries, and therefore are weakly correlated with gender-pronouns in any case. This is of course a hypothesis, and it needs to be explored further.

One important factor to keep in mind when using probes of various grammatical tense, is the context in which they tend to occur. A past tense probe might reflect something that is known and describes a state that has occurred, while a future tense probe might describe potential states. This can affect our analysis as one would expect less discussions about potential occupations for males (assuming that males have access to all) and more mentions about occupations for females (assuming that they have been blocked from male dominated occupations before). This goes back to how genders and occupations are correlated in the training data of pre-trained models, and to what extent this can be perceived when probing the models.

4 Conclusion

We have presented our investigations into how the addition of negation and changing the grammatical tense of the verb in bias probes can alter the correlations between occupations and gendered-pronouns. We carried out experiments using eight pre-trained language models, four Norwegian and four English ones, and generated a set of 16 bias probes.

We show that negation does not have a significant effect on the correlations resulting from probing the language models. However, interesting observations were made for grammatical tense. Switching from present to past shows more correlations with male-gendered pronouns, while changing from present to future exhibits more correlations with female-gendered pronouns. This shows how template-based bias probes are sensitive to small changes, and might hint to the necessity of taking grammatical tense into consideration when probing language models for bias. We believe that aggregating results across tenses might give a better representation of the correlations between genders and occupations.

As future work, we would like to explore the diachronic gender-based bias correlations with occupations. Biases might change across time-periods, and what was not considered bias against one gender a couple of decades ago might now be a stereotypical description. We think that comparing time-periods to each other might help us identify the time-shifts for stereotypical correlations, both in datasets and how this can be reflected in models trained on them.

Acknowledgements

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 Centres 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, 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.

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.

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).

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.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. 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)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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.

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

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.

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

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

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational biases in Norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211, Seattle, Washington. 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.

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.

A Appendix

	Total shift	Shifted to F	Shifted to M	Total shift	Shifted to F	Shifted to M
	jobber som|jobber ikke som			***jobbet som|jobbet ikke som***		
NorBERT	20.39%	93.05%	6.94%	23.51%	100%	0%
NorBERT2	57.50%	0%	100%	39.37%	0%	100%
NB-BERT	25.49%	100%	0%	18.41%	100%	0%
NB-BERT_Large	9.63%	47.05%	52.94%	11.61%	9.75%	90.24%
	skal jobbe som|skal ikke jobbe som			***kommer til å jobbe som|kommer ikke til å jobbe som***		
NorBERT	13.88%	89.79%	10.20%	7.64%	81.48%	18.51%
NorBERT2	41.64%	0%	100%	13.88%	0%	100%
NB-BERT	30.02%	99.05%	0.94%	24.36%	100%	0%
NB-BERT_Large	14.44%	92.15%	7.84%	9.91%	85.71%	14.28%

Table 3: Percentages of occupations that shifted correlations from one gender to another, by adding negations to the Norwegian bias probes.

	Total shift	Shifted to F	Shifted to M	Total shift	Shifted to F	Shifted to M
	works as|does not work as			***worked as|did not work as***		
BERT	6.62%	4.34%	95.65%	3.74%	7.69%	92.30%
BERT_Large	14.12%	6.12%	93.87%	2.30%	12.5%	87.5%
RoBERTa	17.29%	0%	100%	10.66%	2.70%	97.29%
RoBERTa_Large	15.27%	26.41%	73.58%	12.96%	2.22%	97.77%
	will work as|will not work as			***is going to work as|is not going to work as***		
BERT	8.93%	0%	100%	5.76%	0%	100%
BERT_Large	13.25%	2.17%	97.82%	11.81%	2.43%	97.56%
RoBERTa	7.78%	3.70%	96.29%	10.95%	0%	100%
RoBERTa_Large	8.93%	25.80%	74.19%	31.41%	0%	100%

Table 4: Percentages of occupations that shifted correlations from one gender to another, by adding negations to the English bias probes.

comparison	Total shift	shifted to F	Shifted to M	Total shift	shifted to F	Shifted to M
	NorBERT			NorBERT2		
present VS past	14.44%	0%	100%	13.88%	0%	100%
present VS future	11.04%	94.87%	5.12%	12.18%	100%	0%
present VS future2	11.61%	2.43%	97.56%	32.29%	100%	0%
	NB-BERT			NB-BERT_Large		
present VS past	16.43%	0%	100%	7.64%	18.51%	81.48%
present VS future	16.71%	100%	0%	16.43%	96.55%	3.44%
present VS future2	15.01%	100%	0%	13.59%	93.75%	6.25%
	BERT			BERT_Large		
present VS past	5.47%	0%	100%	6.91%	4.16%	95.83%
present VS future	4.03%	35.71%	64.28%	3.74%	30.76%	69.23%
present VS future2	3.74%	15.38%	84.61%	4.89%	52.94%	47.05%
	RoBERTa			RoBERTa_Large		
present VS past	1.15%	0%	100%	15.85%	14.54%	85.45%
present VS future	2.30%	75%	25%	11.81%	39.02%	60.97%
present VS future2	2.30%	0%	100%	27.08%	2.12%	97.87%

Table 5: Total number of occupations that shifted correlations from one gender to another, by changing the tense of the verb in the bias probe. Each tense represents the following probes: present (*jobber ikke som|does not work as a/an*) VS Past (*jobbet ikke som|did not work as a/an*), Future (*skal ikke jobbe som|will not work as a/an*), and Future2 (*kommer ikke til å jobbe som|is not going to work as a/an*).