NorBench – A Benchmark for Norwegian Language Models

Anonymous

Abstract

We present NorBench: a streamlined suite of NLP tasks and probes for evaluating Norwegian language models (LMs) on standardised data splits and evaluation metrics. We also introduce a range of new large-scale Norwegian masked language models with optimised hyperparameters and trained on different text collections. Finally, we compare and analyse their performance, along with other existing Norwegian and multi-lingual LMs, across the different benchmark tests of NorBench.

1 Introduction

The work described in this paper aims to provide a suite of standardised tasks and probes to enable systematic and comparative benchmarking of Norwegian language models (LMs). In addition to collecting a broad range of annotated datasets, we provide precise task definitions, pre-defined data splits and evaluation metrics, with corresponding scripts for streamlining the entire benchmarking pipeline. The resulting resource is dubbed NorBench. We furthermore present a range of new large-scale transformer-based masked language models (LMs) for Norwegian, trained with different parameter configurations and architectures, and on different corpora with different pre-processing. The performance of these models is analysed and compared with respect to NorBench, and also with respect to various dimensions of bias (like gender-bias, toxicity, etc). Our contributions are as follows:

1. We introduce NorBench, a collection of Norwegian datasets and evaluation scripts that ensures simple, fair and standardised comparison between Norwegian LMs. The existing models from prior work are evaluated and compared. All data and code related to NorBench are publicly available online.

Task	Train	Dev	Test
Morpho-syntactic token-level tasks			
UPOS tags	489 217	67 619	54 739
UFeats tags	489 217	67 619	54 739
Lemmas	489 217	67 619	54 739
Dependencies	489 217	67 619	54 739
Named entities	23 071	2 942	2 393
Sentiment analysis			
SA documents	34 903	4 360	4 351
SA sentences	7 973	1 411	1 181
SA targets	5 044	877	735
Linguistic acceptability			
NoCoLA sentences	116 195	14 289	14 383

Table 1: Number of labeled entities in the training, development, and test splits in the datasets used for the NorBench tasks.

2. An integral part of NorBench is diagnostic set of tasks that probe the affinity of LMs towards gender-bias and toxic language – an unfortunate side-effect for many models pre-trained on large amounts of text.

3. We develop a new generation of Norwegian masked LMs (dubbed NorBERT$_3$) that achieves state-of-the-art performance across most NorBench tasks. We provide multiple versions of these models (with different number of parameters) and show that even small versions typically maintain competitive performance.

4. We empirically test the impact of different available Norwegian training corpora on LMs' downstream performance. Our results suggest that pre-training on simple concatenation of all available resources is not always beneficial.

The rest of the paper is structured as follows. Section 2 provides an overview of the tasks in-

cluded in NorBench. In Section 3, the details of our evaluation workflow are outlined. The architecture and the training collections of our novel NorBERT_3 LMs are described in Section 4, while in Section 5, we summarise and analyse the corresponding benchmarking results.Section 6 briefly describes prior work, and in Section 7 we conclude and point out directions for future work.

2 NorBench task descriptions

We here briefly describe each task and associated dataset. The number of training examples for the different datasets and their train, development and test splits are provided in Table 1. For the full details about each task, we refer the reader to the NorBench GitHub repository.[1] Before we describe the various tasks, we first briefly comment on the two official varieties of written Norwegian.

Bokmål and Nynorsk Norwegian has two official written standards: *Bokmål* (BM), used by 85–90% of the Norwegian population, and *Nynorsk* (NN). While they are closely related, there can relatively large lexical differences. The contextualised LMs presented in this paper are therefore trained jointly on both varieties, but with the minority variant Nynorsk represented by comparatively less data than Bokmål (reflecting the natural usage). Several previous studies have indicated that joint modeling of Bokmål and Nynorsk works well for many NLP tasks, like tagging and parsing (Velldal et al., 2017) and NER (Jørgensen et al., 2020). In cases where the labeled data for our benchmark tasks are available as separate versions for Bokmål and Nynorsk, we fine-tune models jointly on the combined BM/NN data. One practical advantage of training joint and 'variety agnostic' models, is that only a single model needs to be maintained, and we bypass the need for a separate 'language identification' step.

2.1 Morpho-syntactic token-level tasks

Part-of-speech (POS) tagging Following Kutuzov et al. (2021), we use the Norwegian Universal Dependencies Treebank (Øvrelid and Hohle, 2016; Velldal et al., 2017) from UD 2.7, in turn converted from NDT (Solberg et al., 2014). The models are challenged to predict universal part-of-speech (UPOS) tags (Nivre et al., 2016). Note that we concatenate the Bokmål (BM) and Nynorsk (NN) datasets (two varieties of Norwegian) for both fine-tuning and evaluation. We use the standard accuracy metric.

Universal feature tagging UPOS tags cover the basic POS categories, but the universal features (UFeats) differentiate more fine-grained lexical and grammatical properties of words. We use the annotations from the same concatenated (BM+NN) UD 2.7 datasets as for the UPOS tags. Similarly to UPOS tagging, UFeats are also evaluated by accuracy.

Lemmatisation Another UD/NDT-based task is lemmatisation, which evaluates how well a language model understands Norwegian morphology to transform an inflected word into its lemmatised form. An integral part of lemmatization – in our variety-agnostic setting – is implicit language identification, because Bokmål and Nynorsk have different standards and the same inflected form sometimes has a different lemma in each language. Correct lemma prediction requires exact match with the gold lemma; we report the aggregated token-wise accuracy.

Dependency parsing The final UD/NDT-based task is dependency parsing, which involves identifying the relationships between words in a sentence, resulting in a dependency tree that represents the grammatical structure of the sentence. By evaluating the quality of dependency parsing outputs by a language model, one can determine its ability to recognise and categorise grammatical roles of words based on their grammatical roles. We report the labelled attachment score (LAS), the standard evaluation metric for dependency parsing,[2], again using joint fine-tuning and testing for BM/NN.

Named entity recognition We use the The NorNE[3] dataset which annotates the UD/NDT (for both Bokmål and Nynorsk) with named entities (Jørgensen et al., 2020). We predict 8 entity types: Person (PER), Organisation (ORG), Location (LOC), Geo-political entity, with a locative sense (GPE-LOC), Geo-political entity, with an organisation sense (GPE-ORG), Product (PROD), Event (EVT), Nominals derived from names (DRV). The evaluation metric used is 'strict' micro F_1, requiring both the correct entity type and exact match of

[1] URL masked for anonymous review.

[2] https://universaldependencies.org/conll18/evaluation.html

[3] https://github.com/ltgoslo/norne

boundary surface string. It is computed using the code for the SemEval'13 Task 9.[4]

2.2 Sentiment analysis

Document-level ternary polarity classification The Norwegian Review Corpus (NoReC; 2nd release) (Velldal et al., 2018) comprises 43 425 professional reviews from a range of Norwegian news sources, and covering a range of different domains (e.g., books, movies, games, music, various consumer goods, etc.). The average length of a document is 389 words. While the reviews originally come with numerical ratings on a scale of 1–6, we here conflate these to three classes; *negative* (ratings 1–3), *fair* (4), and *positive* (5–6). This mapping is done to avoid problems with too few examples for the ratings in the extreme ends of the numerical scale. The dataset comes with predefined datasplits (chronologically sorted), and we evaluate using macro F_1.

Sentence-level ternary sentiment classification We include the dataset $\text{NoReC}_{sentence}$[5] for training and evaluating on the task of sentence-level polarity classification with respect to three classes; (positive, negative or neutral). As described by Kutuzov et al. (2021), this data is derived from NoReC_{fine} (Øvrelid et al., 2020), a subset of NoReC, by aggregating the fine-grained annotations to the sentence-level, removing sentences with mixed sentiment. The evaluation metric is macro F_1.

Targeted sentiment analysis We use the NoReC_{tsa}[6] dataset for the task of targeted sentiment analysis (TSA). As described in Rønningstad et al. (2022), the data is derived from NoReC_{fine} by only including target expressions and the associated positive/negative polarity. The task is to jointly predict the target spans and their polarity, and we use the same evaluation strategy as for NER.

2.3 Linguistic acceptance

NoCoLA The tasks in the Norwegian corpus of linguistic acceptance (NoCoLA; Anonymous, 2023) evaluate language models on their understanding of Norwegian grammaticality. NoCoLA is derived from the ASK Corpus – a language learner corpus of Norwegian as a second language (Tenfjord et al., 2006), which contains texts written exclusively in Norwegian Bokmål, not covering the Nynorsk variety. We report the Matthews correlation coefficient (MCC; Matthews, 1975) on NoCoLA_{class}, the official binary sentence classification variant of the dataset.

2.4 Diagnostics of harmful predictions

Probing for gender-bias and harmful completions Unlike the previous items, this is not a 'task', but rather a description of important model properties. We follow previous works on Norwegian to probe our language models for gender bias in occupations, as well as assessing the harmfulness of their sentence completions (Touileb et al., 2022; Touileb and Nozza, 2022).

3 NorBench baseline methodology

Below we describe various choices pertaining to classifier design when fine-tuning the LMs for the various benchmark tasks. Note that, all of the approaches described here should be considered baselines, in the sense that the goal is not produce state-of-the-art results, but rather to implement simple classification approaches allowing for a fair comparison of different LMs across the various tasks.

3.1 A joint morpho-syntactic model

Since all of these tasks are annotated in the NorNE dataset, we evaluate them jointly with a single multi-task model. We follow the multi-task setup from UDify (Kondratyuk and Straka, 2019) and extend it with BIO label encoding for named-entity recognition. First, we take a separate weighted convex combination of hidden layers for every subtask. Then, we average-pool these contextualised subword representations to get a vector embedding for each token. Finally, these vectors are input to classification heads for UPOS and UFeats tagging, to a classification head for predicting lemma transformation rules, and to biaffine attention heads for dependency parsing (Dozat and Manning, 2017).

3.2 Text classification

Document- and sentence-level sentiment analysis, together with classification of linguistic acceptability, utilise the same text classification approach, based the widely-used fine-tuning scheme from Devlin et al. (2019). There, every tokenised text sequence prefixed by a special `[CLS]` token, ap-

[4] https://github.com/davidsbatista/NER-Evaluation

[5] https://github.com/ltgoslo/norec_sentence

[6] https://github.com/ltgoslo/norec_tsa

pended by a [SEP] token and passed into a pre-trained language model, which produces a contextualised representation for the special [CLS] token. Finally, this representation is passed into the downstream classifier that produces the final prediction among the available classes.

Truncation Document-level sentiment analysis was complicated by the fact that positive classes were not only the largest classes but also had a higher average text length with overall 22% of train set texts being longer than 512 sub-words and a maximum text length of 3 943 sub-words. Several feature engineering strategies were attempted for document-level sentiment analysis, but the most straightforward approach proved the most effective: to simply truncate all texts to the first 512 sub-words. Such a truncation is used for all sequence classification tasks.

3.3 Targeted sentiment analysis

The TSA task is approached as a sequence labeling task, similarly to NER and POS. The models are fine-tuned for 8 epochs, with a training batch size of 32 and learning rate of 5e-05. The best model checkpoint is saved during fine-tuning and applied to make predictions on the test data.

3.4 Probing for gender-bias and harmfulness

We take advantage of the MLM objective of the models, and create templates consisting of gendered head-words, followed by predicates.

Gender-bias To probe for gender bias in occupations, we follow the works of Touileb et al. (2022), and use their templates. These templates are sequences of masked gendered head-words (*e.g.* the woman, the man, the sisters, the uncles ...), followed by predicates pertaining to verbs related to performing an occupation (*e.g.* is, was, worked as, ...), followed by a set of occupations extracted from the Norwegian Statistics bureau (Touileb et al., 2022). Using the probabilities of the masked gendered head-words, we compute the two bias scores: normative and descriptive. The normative score compares the gender-based aggregated probability distributions of templates with a normative distribution of genders in occupations. The idea here is that genders should be equally represented in each occupation with a gender distribution falling between 45% and 55% for each occupation. The descriptive bias score compares the probability distribution of genders across occupations as represented in language models, to the real world distribution of these genders based on the Norwegian Statistics bureau data.

Harmfulness We also follow Touileb and Nozza (2022) to assess the harmfulness of sentence-completions of each language model. We use their templates, constructed similarly to the previous templates where the head-words are gendered-nouns followed by predicates as defined by Nozza et al. (2021), and where the last token is masked. The probing is therefore aiming at completing sentences, by looking at top one, five, ten, and twenty most likely words for each template. Once the completions returned by the models, we compute the HONEST score (Nozza et al., 2021), which is a word-level completion score that maps generated completions to their respective language-specific HurtLex (Bassignana et al., 2018) lexicon of offensive words. The scores represent the total number of completions existing in the lexicon compared to the total amount of returned completions.

4 New Norwegian language models

A number of large Norwegian language models have appeared in recent years: to name only the masked LMs, Kutuzov et al. (2021) trained NorBERT_1, followed by NorBERT_2, and Kummervold et al. (2021) introduced NB-BERT models, coming in different sizes. In this paper, we present a set of novel masked LMs for Norwegian trained with an optimised set of hyperparameters (see below) and on different text collections. We dub these models 'NorBERT_3' and evaluate their performance across different model sizes and training corpora.

4.1 Training corpora

Text sources Our source training dataset included the following text collections:

- Norwegian Wikipedia dumps (BM/NN) from October 2022; about 180 million words;
- NBDigital, public domain texts released by the National Library (NB) of Norway in 2015; 660 million words;[7]
- Norwegian News Corpus (NAK): a collection of Norwegian news texts (both Bokmål and

[7] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-34/

Hyperparameter	x-small	small	base	large
Number of parameters	15M	40M	123M	353M
Number of layers	12	12	12	24
Hidden dimension	192	384	768	1 024
Attention heads	3	6	12	16

Table 2: The main hyperparameters of our four configurations of NorBERT$_3$ language models. Full list of hyperparameters is given in Table 8.

Nynorsk) published between 1998 and 2019; 1.7 billion words;[8]

- Norwegian Colossal Corpus (NCC) the public part of the large and heterogenous corpus released by NB in 2022[9] (Kummervold et al., 2021); about 6.9 billion words;
- Norwegian part of web-crawled mC4 corpus (Xue et al., 2021); about 15 billion words.

The 'standard' NorBERT$_3$ models were trained on the concatenation of these corpora, yielding a training collection of about 25 billion word tokens. In Section 5.2, we investigate the effects of 'oversampling' higher-quality sources and training separate models from scratch on NAK, NCC, mC4, Wikipedia and NBDigital.

Deduplication Before training, all the corpora were de-duplicated on the paragraph-level, using SimHash[10] and removing exact duplicates. The same was done across corpora, reducing their size up to 10%, depending on the corpus.

Filtering Since the largest portion of our training corpus is sourced from web-crawled texts, it is crucial to filter out any unnatural language. Even though our main web-text source is the multilingual Colossal *Clean* Crawled Corpus (mC4), it still contains noisy texts (Dodge et al., 2021), which was also apparent when we manually investigated some of the Norwegian samples. We follow the filtering heuristics implemented for the MassiveText corpus (Rae et al., 2021) and adapt them for Norwegian.

4.2 Architecture

We employ the masked language modeling approach for pre-training NorBERT$_3$ language models. The training roughly follows the BERT recipe (Devlin et al., 2019) with some notable differences:

1. Liu et al. (2019) found out that BERT is undertrained and the next-sentence prediction task is unnecessary – we thus pre-train for $8\times$ more steps, use sequence length of 512 throughout the whole training and utilise only the masked language modeling task (MLM).
2. SpanBERT (Joshi et al., 2020) and T5 (Raffel et al., 2020) demonstrated the advantages of masking random spans instead of individual subwords as in Devlin et al. (2019). Thus, for our MLM objective, the data loader iteratively samples random spans until 15% of the input text is masked. The length of each span is sampled from Geo(p), where $p = 1/3$.
3. In another paper (in print now), we compared various configurations of transformer architectures and of the training hyperparameters. We employ the best performing setting for our pre-training. Crucial upgrades involve using the NormFormer layer normalization (Shleifer and Ott, 2022), disentangled attention with relative positions (He et al., 2021) and increased amount of weight decay.

Parameter count We train four LMs of different sizes (Table 2), accommodating users with varying degrees of computational resources, and to establish a baseline performance across LMs with different number of parameters.

Vocabulary and tokenizer We utilise WordPiece subword tokenizer (Wu et al., 2016) and set its vocabulary size to 50 000. Following GPT-2 (Radford et al., 2019), we represent the text as a sequence of UTF-8 bytes instead of Unicode characters, which substantially reduces the number of out-of-vocabulary tokens. We train the tokenizer on the full corpus utilising the open implementation from the `tokenizers` library.[11]

4.3 Pre-training details

In order to reduce training time, pre-training is paralellised over multiple GPUs with the global batch size of 8 192. The number of GPUs used depends on the size of pre-trained language model, ranging between 16 and 512 AMD Instinct MI250X GPUs, each with 128GB memory. The amount of training

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

[9] https://huggingface.co/datasets/NbAiLab/NCC

[10] https://github.com/ChenghaoMou/text-dedup

[11] https://github.com/huggingface/tokenizers

Model	Size	UPOS	UFeats	Lemma	LAS	NER	Doc. SA	Sent. SA	TSA	NoCoLA
$\text{NorBERT}_{3,\text{ x-small}}$	15M	$98.5^{\pm0.0}$	$97.0^{\pm0.0}$	$97.5^{\pm0.0}$	$92.5^{\pm0.0}$	$84.2^{\pm0.2}$	$64.4^{\pm2.5}$	$66.5^{\pm0.7}$	$25.7^{\pm0.9}$	$47.1^{\pm0.5}$
$\text{NorBERT}_{3,\text{ small}}$	40M	$98.8^{\pm0.0}$	$98.0^{\pm0.0}$	$98.4^{\pm0.0}$	$93.9^{\pm0.1}$	$88.9^{\pm0.3}$	$72.5^{\pm0.5}$	$72.3^{\pm0.4}$	$31.6^{\pm1.1}$	$55.9^{\pm0.2}$
NorBERT_1	111M	$98.6^{\pm0.1}$	$97.8^{\pm0.0}$	$98.5^{\pm0.0}$	$93.3^{\pm0.1}$	$87.1^{\pm0.4}$	$71.4^{\pm1.4}$	$70.6^{\pm0.9}$	$40.7^{\pm1.9}$	$35.9^{\pm1.7}$
NorBERT_2	125M	$98.5^{\pm0.0}$	$97.8^{\pm0.0}$	$98.2^{\pm0.0}$	$93.5^{\pm0.1}$	$85.1^{\pm0.3}$	$72.4^{\pm0.3}$	$73.1^{\pm0.5}$	$39.0^{\pm4.0}$	$56.1^{\pm0.3}$
$\text{NB-BERT}_{\text{base}}$	178M	$98.8^{\pm0.0}$	$98.3^{\pm0.0}$	$98.8^{\pm0.0}$	$94.2^{\pm0.1}$	$89.9^{\pm0.5}$	$73.9^{\pm0.5}$	$74.0^{\pm0.5}$	$45.5^{\pm3.0}$	$58.1^{\pm0.5}$
ScandiBERT	124M	$98.8^{\pm0.0}$	$98.1^{\pm0.0}$	$98.7^{\pm0.0}$	$94.2^{\pm0.1}$	$89.2^{\pm0.3}$	$72.9^{\pm0.4}$	$73.0^{\pm0.6}$	$43.0^{\pm5.5}$	$57.1^{\pm0.4}$
mBERT	178M	$98.4^{\pm0.0}$	$97.4^{\pm0.0}$	$98.3^{\pm0.0}$	$92.6^{\pm0.1}$	$83.9^{\pm0.7}$	$67.4^{\pm0.6}$	$63.5^{\pm1.8}$	$27.7^{\pm7.3}$	$46.4^{\pm0.7}$
$\text{XLM-R}_{\text{base}}$	278M	$98.6^{\pm0.0}$	$97.7^{\pm0.0}$	$98.5^{\pm0.0}$	$93.7^{\pm0.1}$	$87.9^{\pm0.3}$	$71.5^{\pm0.6}$	$72.5^{\pm0.6}$	$43.8^{\pm4.5}$	$58.6^{\pm0.3}$
$\text{NorBERT}_{3,\text{ base}}$	123M	$98.8^{\pm0.0}$	$98.2^{\pm0.1}$	$98.8^{\pm0.0}$	$94.3^{\pm0.1}$	$89.8^{\pm0.4}$	$76.5^{\pm0.2}$	$74.6^{\pm1.3}$	$34.6^{\pm1.0}$	$59.2^{\pm0.3}$
$\text{XLM-R}_{\text{large}}$	560M	$98.6^{\pm0.1}$	$97.8^{\pm0.2}$	$98.5^{\pm0.2}$	$93.9^{\pm0.3}$	$89.0^{\pm0.9}$	$76.5^{\pm0.5}$	$75.6^{\pm1.1}$	$48.1^{\pm1.1}$	$58.6^{\pm0.3}$
$\text{NB-BERT}_{\text{large}}$	355M	$98.5^{\pm0.1}$	$98.1^{\pm0.0}$	$98.3^{\pm0.0}$	$\mathbf{94.5}^{\pm0.1}$	$89.3^{\pm0.1}$	$\mathbf{79.5}^{\pm0.2}$	$77.6^{\pm0.3}$	$\mathbf{50.1}^{\pm3.1}$	$59.7^{\pm0.1}$
$\text{NorBERT}_{3,\text{ large}}$	353M	$\mathbf{98.9}^{\pm0.0}$	$\mathbf{98.4}^{\pm0.0}$	$\mathbf{98.9}^{\pm0.0}$	$\mathbf{94.5}^{\pm0.1}$	$\mathbf{91.3}^{\pm0.3}$	$\mathbf{79.5}^{\pm0.1}$	$\mathbf{78.5}^{\pm1.0}$	$36.0^{\pm0.6}$	$\mathbf{60.9}^{\pm0.1}$

Table 3: NorBench scores for the existing Norwegian language models and our novel NorBERT_3 family of models. We report the mean and standard deviation statistics over 5 runs; the best results are printed in boldface. The 'Size' column reports the number of parameters in the model.

steps is 250 000, increasing the training budget of the original BERT models 8 times.

5 Benchmarking results

In addition to our NorBERT_3 models, we also benchmark these existing Norwegian models:

- **mBERT** (Devlin et al., 2019): multilingual BERT pre-trained on 104 languages, including Norwegian. The training was done exclusively on Wikipedia dumps with oversampled texts from lower resource languages.
- **XLM-R** (Conneau et al., 2020): more advanced multilingual LM that outperformed mBERT on most tasks. XLM-R models were trained on CommonCrawl data for 100 languages.
- **NB-BERT** (Kummervold et al., 2021): $\text{NB-BERT}_{\text{base}}$ model utilised a warm start from pre-trained mBERT. It was later followed by $\text{NB-BERT}_{\text{large}}$ trained from scratch on Norwegian data. Both models are trained on the full – i.e., partially non-public – NCC corpus.
- **NorBERT_1 and NorBERT_2** (Kutuzov et al., 2021): both models follow the pre-training approach of the original BERT model (Devlin et al., 2019). NorBERT_1 is pre-trained on NAK and dumps from both Norwegian Wikipedias, and NorBERT_2 utilises the Norwegian part of mC4 and the public part of NCC.
- **ScandiBERT**: Scandinavian BERT trained on a combination of Danish, Faroese, Icelandic, Norwegian and Swedish texts. However, more than 60% of the training corpus consists of texts from the Norwegian NCC.[12]

5.1 Comparison of models

Table 3 shows results across all the current NorBench tasks for all language models described above (sorted by their size in the number of parameters). Note that we deliberately do not report any average score across all tasks, since we believe that such aggregated scores do not contribute to real understanding of strong and weak sides of different models: one should pay attention to the performance in particular tasks of interest.

Not surprisingly, one can see that it is the largest monolingual models that tend to perform best across the board, and $\text{NorBERT}_{3,\text{ large}}$ (with 353M parameters) specifically obtains the highest scores for most of the tasks except targeted sentiment analysis. At the same time, we see that the smaller models are still very competitive – perhaps most notably $\text{NorBERT}_{3,\text{ small}}$ (with 40M parameters) – and there is certainly an aspect of diminishing returns with increasing the number of parameters.

Gender-bias evaluation Table 4 shows the normative and descriptive occupational bias scores for each model. All models have higher descriptive scores compared to the normative ones, which comes as no surprise. Descriptive scores show how well the models align with the real world distribution of occupations between genders. While no

[12] Short description of its training is available online: `https://huggingface.co/vesteinn/ScandiBERT`.

Model	Normative	Descriptive
NorBERT$_{3,\text{ x-small}}$	**19.78**	37.36
NorBERT$_{3,\text{ small}}$	8.54	34.92
NorBERT$_1$	16.23	39.31
NorBERT$_2$	3.17	34.67
NB-BERT$_{\text{base}}$	18.55	36.50
NorBERT$_{3,\text{ base}}$	13.55	39.43
NB-BERT$_{\text{large}}$	11.35	40.90
NorBERT$_{3,\text{ large}}$	13.67	42.73
NorBERT$_{3,\text{ base}}$, oversampled	9.64	36.99
NorBERT$_{3,\text{ base}}$, NAK only	14.04	49.81
NorBERT$_{3,\text{ base}}$, NCC only	12.57	48.84
NorBERT$_{3,\text{ base}}$, mC4 only	11.72	39.31
NorBERT$_{3,\text{ base}}$, NB only	12.33	38.21
NorBERT$_{3,\text{ base}}$, Wiki only	15.99	**50.42**

Table 4: Normative and descriptive occupational bias scores. Best scores in bold, and worst scores underlined.

Model	k = 1	k = 5	k = 10	k = 20
NorBERT$_{3,\text{ x-small}}$	0.0062	0.0062	0.0040	0.0037
NorBERT$_{3,\text{ small}}$	0.0015	0.0018	0.0027	0.0049
NorBERT$_1$	0.0310	**0.0378**	**0.0306**	**0.0258**
NorBERT$_2$	**0.0356**	0.0229	0.0189	0.0159
NB-BERT$_{\text{base}}$	0.0124	0.0083	0.0080	0.0069
NorBERT$_{3,\text{ base}}$	0.0	0.0027	0.0026	0.0055
NB-BERT$_{\text{large}}$	0.0294	0.0285	0.0279	0.0244
NorBERT$_{3,\text{ large}}$	0.0	0.0006	0.0013	0.0033
NorBERT$_{3,\text{ base}}$, oversampled	0.0046	0.0071	0.0085	0.0092
NorBERT$_{3,\text{ base}}$, NAK only	0.0093	0.0080	0.0093	0.0125
NorBERT$_{3,\text{ base}}$, NCC only	0.0	0.0006	0.0010	0.0028
NorBERT$_{3,\text{ base}}$, mC4 only	0.0	0.0003	0.0009	0.0038
NorBERT$_{3,\text{ base}}$, NB only	0.0015	0.0031	0.0012	0.0026
NorBERT$_{3,\text{ base}}$, Wiki only	0.0	0.0012	0.0071	0.0082

Table 5: The harmfulness score of models looking at top one, five, ten, and twenty most likely completions using HONEST (Nozza et al., 2021).

model achieves a perfect score, the top three best models are the NorBERT$_{3,\text{ base}}$ trained on respectively Wikipedia, NAK, and NCC. The nature of these corpora leads to increased correlations between gendered-nouns and occupations, as they usually tend to be described in a descriptive way. NorBERT$_{3,\text{ base, oversampled}}$ achieves the worst descriptive bias score of all models, but still scoring better than the best model on the normative score. Looking more specifically at gender-dominated and gender-neutral occupations, it is clear that all models are much better at identifying female-dominated occupations. All models achieve very low scores on gender-neutral occupations, suggesting a tendency to correlate occupations with one gender, rather then equally representing them. These results can be seen in Table 7 in the Appendix.

On the other hand, where we expect genders to be equally represented, no model achieves as high scores in the normative scores, as in the descriptive ones. The best model is the smallest model NorBERT$_{3,\text{ x-small}}$, which might suggest that from a normative perspective, the smaller the model, the more balanced representation of genders when it comes to occupations.

Harmfulness scores In addition to the normative and descriptive occupational bias scores, we also compute the harmfulness of the sentences completed by these models. Table 5 shows the HONEST scores (Nozza et al., 2021) of each model. Here we evaluate the top-k completions, where we look at the first, five, ten, and twenty most likely completions. Overall, NorBERT$_3$ models achieve very low harmfulness scores compared to the other Norwegian language models. NorBERT$_{3,\text{ base}}$, trained on mC4 only, achieves the lowest scores across all top k completions, which is quite interesting as this corpus is a subset of the Common Crawl's corpus, which is known to contain problematic content. It is closely followed by NorBERT$_{3,\text{ large}}$. NorBERT$_1$ is the worst model across all top k completions.

5.2 Comparison of Norwegian corpora

The downstream performance of a language model is a result of a combination of training choices and choices of the training corpus. In order to study the second aspect, we fix the training configuration and pre-train multiple NorBERT$_{3,\text{ base}}$ models on different Norwegian corpora.

We compare a simple concatenation of all available resources (*'combined'*) against a variant with oversampling the quality data. The reasoning behind this was that the mC4 corpus is the most noisy of all the above, since it is created by web crawling. We hypothesised that artificially increasing the amount of data from the cleaner corpora (Wikipedia, NBDigital, NCC and NAK) will improve the resulting model's performance. We implemented this by creating an *'oversampled'* train collection where all the sentences from the clean corpora were repeated twice, so that the total size of the 'clean' part approximately matched the size of the mC4 corpus.

In addition, to study the respective usefulness of

Corpus	UPOS	UFeats	Lemma	LAS	NER	Doc. SA	Sent. SA	TSA	NoCoLA
Combined	**98.8**$^{\pm 0.0}$	**98.2**$^{\pm 0.1}$	**98.8**$^{\pm 0.0}$	94.3$^{\pm 0.1}$	89.8$^{\pm 0.4}$	**76.5**$^{\pm 0.2}$	74.6$^{\pm 1.3}$	34.6$^{\pm 1.0}$	**59.2**$^{\pm 0.3}$
Oversampled	**98.8**$^{\pm 0.0}$	98.1$^{\pm 0.0}$	98.7$^{\pm 0.0}$	94.3$^{\pm 0.1}$	89.3$^{\pm 0.2}$	73.6$^{\pm 0.6}$	75.5$^{\pm 0.9}$	34.0$^{\pm 0.4}$	57.6$^{\pm 0.1}$
NAK	**98.8**$^{\pm 0.0}$	98.0$^{\pm 0.0}$	98.5$^{\pm 0.0}$	**94.4**$^{\pm 0.0}$	**90.0**$^{\pm 0.4}$	**76.5**$^{\pm 0.2}$	**77.5**$^{\pm 0.7}$	**35.2**$^{\pm 0.9}$	58.3$^{\pm 0.3}$
NCC	**98.8**$^{\pm 0.0}$	**98.2**$^{\pm 0.0}$	98.7$^{\pm 0.0}$	94.3$^{\pm 0.0}$	89.3$^{\pm 0.3}$	73.6$^{\pm 0.2}$	75.2$^{\pm 0.6}$	32.5$^{\pm 0.4}$	58.3$^{\pm 0.4}$
mC4	**98.8**$^{\pm 0.0}$	**98.2**$^{\pm 0.0}$	98.6$^{\pm 0.0}$	94.3$^{\pm 0.1}$	89.9$^{\pm 0.4}$	74.2$^{\pm 0.6}$	76.8$^{\pm 0.7}$	34.8$^{\pm 0.9}$	58.5$^{\pm 0.3}$
Wiki	98.6$^{\pm 0.0}$	97.7$^{\pm 0.0}$	98.3$^{\pm 0.0}$	93.7$^{\pm 0.1}$	87.9$^{\pm 0.3}$	68.5$^{\pm 0.3}$	68.2$^{\pm 0.7}$	31.1$^{\pm 0.7}$	54.1$^{\pm 0.3}$
NBDigital	**98.8**$^{\pm 0.0}$	98.1$^{\pm 0.0}$	98.7$^{\pm 0.0}$	94.0$^{\pm 0.0}$	87.0$^{\pm 0.4}$	71.1$^{\pm 0.2}$	69.9$^{\pm 0.7}$	30.6$^{\pm 1.0}$	56.1$^{\pm 0.1}$

Table 6: The downstream performance of NorBERT$_{3,\text{ base}}$ models pre-trained on different corpora. We report the mean and standard deviation statistics over 5 runs; the best results are printed in boldface.

particular Norwegian text collections, we trained separate models from scratch on NAK, NCC, mC4, Wikipedia and NBDigital.

Results The NorBench results for the corpora comparison are given in Table 6. We believe there are three noteworthy – and perhaps surprising – take-aways of this comparison:

1. We hypothesised that oversampling the high-quality texts should lead to increased downstream performance. This is evidently a false assumption as oversampling works slightly worse overall. Large language models are known to be sensitive to duplicate data (Lee et al., 2022), which might explain such a behaviour.
2. A straightforward concatenation of all available resources does not necessarily lead to better performance – but it is a reasonable approach for a general model. The simple concatenation works the best on average. On the other hand, pre-training only on NAK leads to substantially improved performance on sentence-level and targeted sentiment analysis, perhaps due to the close match in terms of text type (news articles).
3. Even though Wikipedia must be considered a small resource of natural texts by today's standards, the LM pre-trained only on Norwegian Wikipedia data is surprisingly competitive (especially on morpho-syntactic tasks), even outperforming some of the pre-existing Norwegian models. This demonstrates that it is possible to develop reliable models for low-resource language or specific language domains.

6 Related work

Evaluating pre-trained language models for particular languages and cross-lingually is a venerable research sub-field within NLP. Well-known benchmark sets for English include GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019) and GLGE (Liu et al., 2021), among others. However, up to now benchmarking LMs for Norwegian was limited to separate test sets with non-standardised evaluation workflows. ScandEval (Nielsen, 2022) is an attempt to create a standard natural language understanding benchmark across Scandinavian languages (Danish, Swedish and Norwegian). However, it is not well-described in academic publications and, even more importantly, does not focus on evaluating specifically Norwegian tasks (only four of the test sets are Norwegian). We address this issue with NorBench.

7 Summary

We presented NorBench, a set of standardised benchmark sets for evaluating Norwegian language models. The aim of this effort is to provide NLP practitioners with a comprehensive and streamlined service, including a leaderboard, human-annotated datasets, evaluation workflow, and open code implementing this workflow.

This paper also describes and evaluates a set of novel NorBERT$_3$ masked LMs trained on several different Norwegian text collections in different model sizes. They are shown to outperform Norwegian LMs from prior work on the majority of NorBench tasks.

We consider NorBench to be a dynamic resource that we plan to continually extend in future work, to support additional tasks and additional architectures. While we anticipate including new annotated benchmark data as they may become available in the future, there are also existing datasets that we plan to include in the shorter term, like coreference resolution based on the NARC dataset (Mæhlum et al., 2022), negation resolution based on NoReC$_{neg}$ (Mæhlum et al., 2021) and NorQuAD, an upcoming question-answering dataset.

References

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the 5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

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.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep bi-affine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. NorNE: Annotating Named Entities for Norwegian. In *Proceedings of the 12th Edition of the Language Resources and Evaluation Conference*, Marseille, France, 2020.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Per E 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 *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for Norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou, Ming Gong, Pengcheng Wang, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, Ruofei Zhang, Winnie Wu, Ming Zhou, and Nan Duan. 2021. GLGE: A new general language generation evaluation benchmark. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 408–420, Online. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Petter Mæhlum, Jeremy Barnes, Robin Kurtz, Lilja Øvrelid, and Erik Velldal. 2021. Negation in Norwegian: an annotated dataset. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics*.

Petter Mæhlum, Dag Haug, Tollef Jørgensen, Andre Kåsen, Anders Nøklestad, Egil Rønningstad, Per Erik Solberg, Erik Velldal, and Lilja Øvrelid. 2022. Narc–norwegian anaphora resolution corpus. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 48–60.

B.W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Dan Saattrup Nielsen. 2022. Scandeval: Evaluation of language models on mono- or multilingual scandinavian language tasks. *GitHub. Note: https://github.com/saattrupdan/ScandEval*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

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*, pages 2398–2406, Online. Association for Computational Linguistics.

Lilja Øvrelid and Petter Hohle. 2016. Universal Dependencies for Norwegian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 1579–1585, Portorož, Slovenia.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the 12th Edition of the Language Resources and Evaluation Conference*, Marseille, France.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *CoRR*, abs/2112.11446.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2022. Entity-level sentiment analysis (ELSA): An exploratory task survey. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6773–6783, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sam Shleifer and Myle Ott. 2022. Normformer: Improved transformer pretraining with extra normalization.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The Norwegian Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Kari Tenfjord, Paul Meurer, and Knut Hofland. 2006. The ASK corpus - a language learner corpus of Norwegian as a second language. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Samia Touileb and Debora Nozza. 2022. Measuring harmful representations in Scandinavian language models. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 118–125, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian Review Corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, pages 4186–4191, Miyazaki, Japan.

Erik Velldal, Lilja Øvrelid, and Petter Hohle. 2017. Joint UD parsing of Norwegian Bokmål and Nynorsk. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, pages 1–10, Gothenburg, Sweden.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

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.

8 Appendix

Model	N	F	M
NorBERT$_{3,\text{ x-small}}$	2.19	31.01	4.15
NorBERT$_{3,\text{ small}}$	0.48	33.69	0.73
NorBERT$_{3,\text{ base}}$	1.22	33.21	5.00
NorBERT$_{3,\text{ large}}$	1.70	33.33	7.69
NorBERT$_{3,\text{ base}}$, oversampled	0.36	33.45	3.17
NorBERT$_{3,\text{ base}}$, NAK only	2.56	28.81	18.43
NorBERT$_{3,\text{ base}}$, NCC only	2.19	30.76	15.87
NorBERT$_{3,\text{ base}}$, mC4 only	0.61	33.33	5.37
NorBERT$_{3,\text{ base}}$, NB only	0.73	30.03	7.44
NorBERT$_{3,\text{ base}}$, Wiki only	2.56	25.88	21.97

Table 7: Descriptive bias scores of gender-dominated and gender-neutral occupations. Where N stands for neutral, F for female, and M for male.

Hyperparameter	**NorBERT$_{3,\text{ x-small / small / base / large}}$**
Number of layers	12 / 12 / 12 / 12
Hidden size	192 / 384 / 768 / 1 024
FF intermediate size	2 048
Vocabulary size	16 384
FF activation function	GEGLU
Attention heads	12
Attention head size	64
Dropout	0.1
Attention dropout	0.1
Training steps	250 000
Batch size	8 192
Sequence length	512
Warmup steps	4 000 (1.6% steps)
Initial learning rate	0.01
Final learning rate	0.001
Learning rate decay	cosine
Weight decay	0.1
Layer norm ϵ	1e-5
Optimizer	LAMB
LAMB ϵ	1e-6
LAMB β_1	0.9
LAMB β_2	0.98
Gradient clipping	2.0

Table 8: Pre-training hyperparameters. The models differ only in their hidden size and number of layers, the learning rate schedule and other training settings are kept identical.