SocioProbe: What, When, and Where Language Models Learn about Sociodemographics

Anne Lauscher[1], Federico Bianchi[2], Samuel Bowman[3], and Dirk Hovy[4]
[1]Data Science Group, University of Hamburg, Germany
[2]StanfordNLP, Stanford University, CA, USA
[3]New York University, NY, USA
[4]MilaNLP, Bocconi University, Milan, Italy
anne.lauscher@uni-hamburg.de, fede@stanford.edu,
bowman@nyu.edu, dirk.hovy@unibocconi.it

Abstract

Pre-trained language models (PLMs) have outperformed other NLP models on a wide range of tasks. Opting for a more thorough understanding of their capabilities and inner workings, researchers have established the extend to which they capture lower-level knowledge like grammaticality, and mid-level semantic knowledge like factual understanding. However, there is still little understanding of their knowledge of higher-level aspects of language. In particular, despite the importance of sociodemographic aspects in shaping our language, the questions of whether, where, and how PLMs encode these aspects, e.g., gender or age, is still unexplored. We address this research gap by probing the sociodemographic knowledge of different single-GPU PLMs on multiple English data sets via traditional classifier probing and information-theoretic minimum description length probing. Our results show that PLMs do encode these sociodemographics, and that this knowledge is sometimes spread across the layers of some of the tested PLMs. We further conduct a multilingual analysis and investigate the effect of supplementary training to further explore to what extent, where, and with what amount of pre-training data the knowledge is encoded. Our overall results indicate that sociodemographic knowledge is still a major challenge for NLP. PLMs require large amounts of pre-training data to acquire the knowledge and models that excel in general language understanding do not seem to own more knowledge about these aspects.

1 Introduction

When talking to somebody, we consciously choose how to represent ourselves, and we have a mental model of who our conversational partner is. At the same time, our language is littered with subconscious clues about our sociodemographic background that we cannot control (e.g., our age, education, regional origin, social class, etc). People use this information as an integral part of language, to better reach their audience, and to understand what they are saying (e.g., Trudgill, 2000). In other words, we use sociodemographic knowledge to decide what to say (are we talking to a child or an adult, do I want to sound smart or relatable?) But do pre-trained language models (PLMs) have knowledge about sociodemographics?

Over the last years, PLMs like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have achieved superior performance on a wide range of downstream tasks (e.g., Wang et al., 2018, 2019, *inter alia*). Accordingly, they have become the *de facto* standard for most NLP tasks. Consequently, many researchers have tried to shed light on PLMs' inner workings (cf. *"Bertology"*; Tenney et al., 2019; Rogers et al., 2020). They have systematically probed the models' capabilities to unveil which language aspects their internal representations capture. In particular, researchers have probed lower-level structural knowledge (e.g., Hewitt and Manning, 2019; Sorodoc et al., 2020; Chi et al., 2020; Pimentel et al., 2020, *inter alia*), as well as mid-level knowledge, e.g., lexico-semantic knowledge (e.g., Vulić et al., 2020; Beloucif and Biemann, 2021), and PLMs' factual understanding (e.g., Petroni et al., 2019; Zhong et al., 2021). While these aspects are relatively well explored, we still know little about higher-level knowledge of PLMs: only a few works have attempted to quantify common sense knowledge in the models (Petroni et al., 2019; Lin et al., 2020). Probing of other higher-level aspects still remains underexplored – hindering targeted progress in advancing human-like natural language understanding.

As recently pointed out by Hovy and Yang (2021), sociodemographic aspects play a central role in language. However, they remain underexplored in NLP, despite promising initial findings (e.g., Volkova et al., 2013; Hovy, 2015; Lynn et al., 2017). Importantly, we are not aware of *any* research assessing sociodemographic knowledge in

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7901–7918
December 7-11, 2022

PLMs. This lack is extremely surprising given the availability of resources, and the importance of these factors in truly understanding language.

Contributions. Acknowledging the importance of sociodemographic factors in language, we address a research gap by proposing SOCIOPROBE, a novel perspective of probing PLMs for sociodemographic aspects. We demonstrate our approach along two dimensions, (binary) *gender* and *age*, using two established data sets, and with different widely-used easily-downloadable PLMs that can be run on a single GPU. To ensure validity of our findings, we combine "traditional" classifier probing (Petroni et al., 2019) and information-theoretic minimum distance length (MDL) probing (Voita and Titov, 2020). Our experiments allow us to answer a series of research questions. We find that PLMs *do* represent sociodemographic knowledge, but that it is acquired in the later stages. This knowledge is also decoupled from overall performance: some models that excel in general language understanding do still not have more knowledge about sociodemographics encoded. We hope that this work inspires more research on the social aspects of NLP. Our research code is publicly available at https://github.com/MilaNLProc/socio-probe.

2 Research Questions

We pose five research questions (**RQs**):

RQ1: ***To what extend do current PLMs encode sociodemographic knowledge?*** Do these models "know" about the existence and impact of sociodemographic aspects like age or gender on downstream tasks, as repeatedly shown (e.g., Volkova et al., 2013; Hovy, 2015; Benton et al., 2017)? We probe different versions of the RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021b,a) model families. Our findings reveal the varying extent to which sociodemographic knowledge is encoded in different textual domains. Surprisingly, the superior performance of the DeBERTa model on general NLU tasks is not reflected in the encoding of sociodemographic knowledge.

RQ2: ***How much pre-training data is needed to acquire sociodemographic knowledge?*** Are sociodemographic aspects present in any data sample, or are they only learned with sufficient amounts of data? Inspired by Zhang et al. (2021b), we use a suite of MiniBERTas (Warstadt et al., 2020) and RoBERTa *base* trained on different amounts of data (1M to 30B). Our results show that sociodemographic knowledge is learned much more slowly than syntactic knowledge and the gains do not seem to flatten with more training data. This indicates that large data portions are needed to acquire sociodemographic knowledge.

RQ3: ***Where is sociodemographic knowledge located in the models?*** Sociodemographic aspects influence a wide range of NLP tasks, both at a grammatical level (e.g., part-of-speech tagging Garimella et al. (2019)) and at a pragmatic level (e.g., machine translation Hovy et al. (2020); Saunders and Byrne (2020)). But where do these factors reside themselves in the model? By probing different layers of the PLM with SOCIOPROBE, we find that sociodemographic knowledge is located in the higher layers of most PLMs. This finding is in-line with the intuition that higher-level semantic knowledge is encoded in higher layers of the models (e.g., Tenney et al., 2019). However, on some data sets, some of the models show an opposite trend and the differences across the layers seem much less pronounced than for a lower-level control task, in which we predict linguistic acceptability.

RQ4: ***Does the localization of sociodemographic knowledge in multilingual models differ?*** Different languages provide different linguistic ways of expressing sociodemographic (and other) aspects: some lexically, some syntactically (Johannsen et al., 2015). Do PLMs that have been exposed to multiple languages store sociodemographic knowledge differently than monolingual models? We probe multilingual models and demonstrate that the results are in-line with the findings from RQ3. Thus, the localization of the sociodemographic knowledge in the multilingual versions does not seem to differ from their monolingual counterparts.

RQ5: ***What is the effect of different supplementary training tasks on the knowledge encoded in the PLMs' features?*** Phang et al. (2018) demonstrated that through supplementary training on intermediate-labeled tasks (STILTs), the performance for downstream tasks can be improved. We hypothesize that such sequential knowledge transfer can activate sociodemographics in PLMs, as these aspects can act as useful signals, e.g., for sentiment analysis (Hovy, 2015). However, our experiments show that specifically the sociodemographic knowledge in the last layers of the models is overwritten through our STILTs procedures.

Overall, the encoding of sociodemographic knowledge is still a major challenge for NLP: **models that excel in NLU do not have more knowledge about sociodemographics, learning curves do not flatten with more pretraining data, the knowledge is much less located than for other tasks, and learning from other tasks is difficult**.

3 Related Work

Probing PLMs. The success of large PLMs has led to researchers developing a range of methods (e.g., Hewitt and Liang, 2019; Torroba Hennigen et al., 2020) and data sets (e.g., Warstadt et al., 2020; Hartmann et al., 2021) for obtaining a better understanding of PLMs. In turn, those approaches also challenge these paradigms (e.g., Pimentel et al., 2020; Ravichander et al., 2021). The most straightforward probing approach relies on training classifiers (e.g., Petroni et al., 2019) to probe models' knowledge. In contrast, other probing mechanisms are substractive (Cao et al., 2021), intrinsic (Torroba Hennigen et al., 2020), or rely on control tasks (Hewitt and Liang, 2019). A popular family is information theoretic probing (e.g., Pimentel et al., 2020; Pimentel and Cotterell, 2021), like minimum description length (MDL) probing (Voita and Titov, 2020). We use MDL complementarily to traditional probing to further substantiate our claims. Most authors have focused on probing English language models (e.g., Conneau et al., 2018; Liu et al., 2021; Wu and Xiong, 2020; Koto et al., 2021, *inter alia*), but some have moved into the multilingual space (e.g., Ravishankar et al., 2019; Kurfalı and Östling, 2021; Shapiro et al., 2021), or probed multimodal models (e.g., Prasad and Jyothi, 2020; Hendricks and Nematzadeh, 2021).

Researchers have used probing to understand whether PLMs encode knowledge about several aspects of language, and to which extent: researchers have probed PLMs for syntactic knowledge (e.g., Hewitt and Manning, 2019; Sorodoc et al., 2020), lexical semantics (Vulić et al., 2020; Beloucif and Biemann, 2021), factual knowledge (Heinzerling and Inui, 2021; Petroni et al., 2019; Zhong et al., 2021), and common sense aspects (Lin et al., 2020) or domain-specific knowledge (Jin et al., 2019; Pandit and Hou, 2021; Wu and Xiong, 2020). Despite this plethora of works, the sociodemographic knowledge remains underexplored.

NLP and Sociodemographic Aspects. Our language use varies depending on the characteristics of the sender and receiver(s), e.g., their age and gender (Eckert and McConnell-Ginet, 2013; Hovy and Yang, 2021). Accordingly, researchers in NLP have explored these variations (Rosenthal and McKeown, 2011; Blodgett et al., 2016) and showed that sociodemographic factors influence model performance (e.g., Volkova et al., 2013; Hovy, 2015). Since then, many researchers have argued that such factors should be taken into account for human-centered NLP (Flek, 2020), and showed gains from sociodemographic adaptation (e.g., Lynn et al., 2017; Yang and Eisenstein, 2017; Li et al., 2018).

Other researchers have exploited the tie between language and demographics to profile authors from their texts (Burger et al., 2011; Nguyen et al., 2014; Ljubešić et al., 2017; Martinc and Pollak, 2018). In this work, we do not develop methods to predict demographic aspects, but use this task as a proxy to how well sociodemographic knowledge is encoded in our models. Another line of research has worked on detecting and removing unfair stereotypical bias towards demographic groups from PLMs (Blodgett et al., 2020; Shah et al., 2020), e.g., gender bias (May et al., 2019; Lauscher and Glavaš, 2019; Webster et al., 2020; Lauscher et al., 2021). Most recently and closest to our work, Zhang et al. (2021a)[1] investigate the sociodemographic bias of PLMs. They compare the PLMs cloze predictions with answers given by crowd workers belonging to different sociodemographic groups. However, they do not provide further insights of the nature of this knowledge nor how when and where it is encoded. Our work unequivocally establishes that PLMs contain sociodemographic knowledge, and shows how it is likely acquired, and where it resides.

4 SocioProbe

We describe SocioProbe, which we employ to explore the sociodemographic knowledge PLMs contain. Guided by the availability of data sets, we focus on the dimensions of *gender* and *age*. Note, however, that our overall methodology can be easily extended to other sociodemographic aspects.

4.1 Data

We probe sociodemographic aspects on two data sets. They vary in terms of text length and domain.

[1]Note that their interest is in linguistically determined language varieties of social groups, i.e., sociolects, whereas we focus on the interplay between individual *demographic* aspects that go across language varieties: we can express gender independent of whether we speak in dialect or standard language.

Dataset Name	Textual Domain	Dimension	Label	# Instances	% Instances
Trustpilot	Product Reviews	Gender	*Man* *Woman*	5349	49.97 50.03
		Age	*Young* *Old*	5269	52.19 47.80
RTGender	Facebook Posts (Congress Members)	Gender	*Man* *Woman*	510135	75.16 24.84
	Facebook Posts (Public Figures)	Gender	*Man* *Woman*	133,017	33.38 66.62
	Fitocracy Posts	Gender	*Man* *Woman*	318,535	54.54 45.46
	Reddit Posts	Gender	*Man* *Woman*	1,453,512	79.02 20,98

Table 1: Datasets with dimensions, number of instances (# Instances), and label distributions (% Instances).

Trustpilot (Hovy, 2015). Trustpilot[2] is an international user review platform. The data consists of the review texts (including the rating, which we do not use in this work), as well as the self-identified gender and age of the author. Following the original paper, we do not consider users from 35 to 45 to reduce possible errors due to noisy boundaries. We use the split introduced by Hovy et al. (2020), and focus on the English portion of the data set. For age, we use *Young* for users under the age of 35, and *Old* for people above the age of 45.

RTGender (Voigt et al., 2018). We use all texts of the data set from three different social media platforms: Reddit,[3] Facebook (posts from politicians and public figures),[4] and Fitocracy.[5] Our true label corresponds to the gender of the author. In total, the data set consists of 2,415,199 instances. For our experiments, we subsample 20,000 samples for each domain to start from equally-sized portions.

4.2 Probing Methodology

We combine two probing methodologies: traditional classifier probing and MDL probing.

Traditional Classifier Probing. The traditional approach to PLM probing is to place a simple classifier – the probe – on top of the frozen features (e.g., Ettinger et al., 2016; Adi et al., 2016, *inter alia*). In our case, following Tenney et al. (2019) and Zhang et al. (2021b), we use a simple two-layer feed-forward network (with rectified linear unit as the activation function) with a softmax output layer. We feed it the average hidden representations of the PLM's Transformer. We take care to only average over the representations of the text and ignore special tokens. We report the F1 measure.

Minimum Description Length Probing. Traditional classifier probing has been criticized for its reliance on the complexity of the probe (Hewitt and Liang, 2019; Voita and Titov, 2020). To ensure validity of our results, we thus combine classifier probing with an information theoretic approach. Concretely, we use MDL (Voita and Titov, 2020). The intuition behind MDL is that the more information is encoded, the less data is needed to describe the labels given the representations. As in the implementation of the *online code estimation setting*, we partition the data into 11 non-overlapping portions representing 0%, 0.1%, 0.2%, 0.4%, 0.8%, 1.6%, 3.2%, 6.25%, 12.5%, 25%, 50%, and 100% of the full data sets with t numbers of examples each: $\{(x_j, y_j)\}_{j=t_{i-1}+1}^{t_i}$ for $1 \leq i \leq 11$. Next, we train a classifier on each portion i and compute the Loss $\mathcal{L}$ on the next portion $i+1$. The codelength corresponds to the sum of the resulting 10 losses plus the codelength of the first data portion:

$$\text{MDL} = t_1 \log_2 2 - \sum_{i=1}^{10} \mathcal{L}_{i+1}\,, \qquad (1)$$

with t_1 as the number of training examples in the first portion of the data set. A lower MDL value indicates more expressive features.

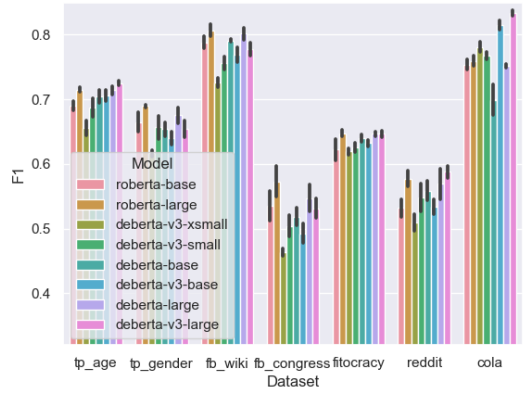
5 Experiments

We describe our experiments.

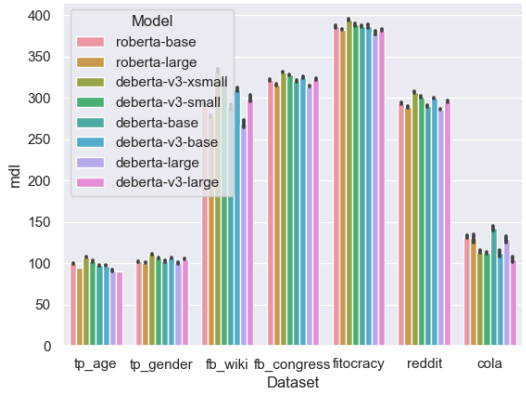
5.1 General Experimental Setup

All our experiments follow roughly the same experimental setup: For the Trustpilot data, we use

[2] https://www.trustpilot.com
[3] https://www.reddit.com
[4] https://www.facebook.com
[5] https://www.fitocracy.com



(a) Classic probing



(b) MDL probing

Figure 1: Results for RQ1. We compare different RoBERTa and DeBERTa models (RoBERTa *base*, RoBERTa *large*, DeBERTa *base*, DeBERTa *large*, DeBERTa v3 *xsmall*, *small*, *base*, and *large*) for (a) classic and (b) MDL probing. We report average and standard deviation of the F1-scores for 5 runs across 7 tasks.

the standard splits provided in Hovy (2015). On all other data sets, described in Section 4.1 we apply a standard split, with 80% of the data for training, 10% for validation, and 10% for testing the models. We train all our models in batches of 32 with a learning rate of 1e-3 using the Adam optimizer (Kingma and Ba, 2015) (using default parameters from pytorch). We apply early stopping based on the validation set loss with a patience of 5 epochs. If the loss does not improve for an epoch, we reduce the learning rate by 50 %. We conduct all experiments 5 times with different random initializations of the probes and report the mean and the standard deviation of the performance scores. For all models, we use versions available on Huggingface and we provide links to all models and code bases used in the Supplementary Materials.

5.2 RQ1: To what extend do PLMs encode sociodemographic knowledge?

As initial base experiment, we want to establish how well sociodemographic knowledge can be predicted from the features of different PLMs.

Approach. We test the features extracted from RoBERTa (Liu et al., 2019) in *base* and *large* configuration in comparison to DeBERTa (He et al., 2021b) in *xsmall*, *small*, *base*, and *large* configuration. For DeBERTa, different versions are available in the Huggingface repository. We use the original model as well as the v3 version (He et al., 2021a) of *base* and *large*. The v3 employs ELECTRA-style pre-training with gradient disentangled embedding sharing (Clark et al., 2020) leading to improvements across all GLUE tasks.[6]

Results. Figure 1 shows the results. Generally, the trends in the different models are consistent across the two different probing approaches (classic probing and MDL probing). Therefore, we conclude the validity of our approach. *The difficulty of the different data sets varies*: the "easiest" task is our control task CoLA, in which we probe lower-level syntactic knowledge. The next-easiest task is to predict the gender in Facebook posts of public figures (*fb_wiki*, e.g., 80.68 % average F1 score for RoBERTa *large*). In contrast, predicting the gender of Facebook posts of congress members is relatively difficult for the models (*fb_congress*, e.g., 57.32 % average F1 score). This is in line with the findings of Voigt et al. (2018): depending on the domain of text, the sociodemographic aspects of authors are reflected to varying degrees (here: less so in more formal settings). Interestingly, we can not confirm the overall superiority of the DeBERTa models. While the DeBERTa v3 *base* and *large* models outperform RoBERTa on CoLA by a large margin (6.93 percentage points difference between RoBERTa large and DeBERTa large, as per He et al., 2021a), *RoBERTa large seems to encode sociodemographic knowledge similarly well as DeBERTa large*, or to an even larger extent. The same observation holds when comparing DeBERTa versions. This finding warrants further investigation into how different training regimes affect the encoding of higher-level knowledge.

[6] https://github.com/microsoft/DeBERTa#fine-tuning-on-nlu-tasks

# Tokens	Costs ($)	CO_2 (lbs)	μGain
1M	50	5.825	–
10M	500	58.250	+2.61
100M	5,075	582.500	+1.98
1B	20,320	2,330.000	+0.30
30B	609,600	69,900.000	+8.56

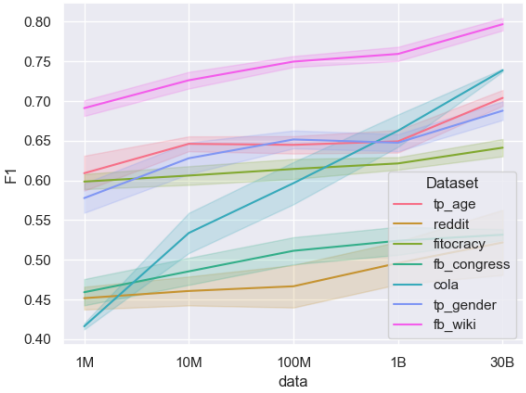
Table 2: Results of our cost-benefit analysis. We show financial costs (Costs ($)) and CO_2 emissions (CO_2 (lbs)), gain is average F1-measure increase over the ext smaller model across all data sets and models (μGain).

5.3 RQ2: How much pre-training data is needed to acquire sociodemographic knowledge?

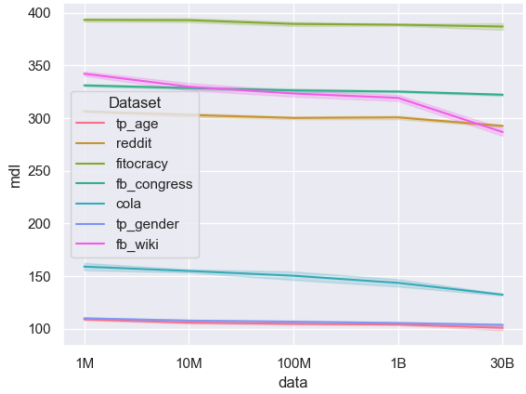
We test models trained on varying amounts of data.

Approach. We use the suite of MiniBERTas (Warstadt et al., 2020), 12 RoBERTa-like models, which have been trained on 1M, 10M, 100M, and 1B words, respectively.[7] The data was randomly sampled from a corpus similar to the original BERT (Devlin et al., 2019). Pretraining data consisted of the English Wikipedia and Smashwords, which is similar to the BookCorpus (Zhu et al., 2015). The size of the model trained on the smallest portion (1M) is *medium small* (6 layers, 8 attention heads, hidden size of 512). The other models were trained with the *base* configuration (12 layers, 12 attention heads, hidden size of 768). For each size, 3 checkpoints are available (the ones which yielded lowest validation perplexity), trained with different hyperparameters. In comparison, we probe the original RoBERTa in *base* configuration (Liu et al., 2019), trained on approximately 30B words.

Results. The results for the classic and MDL probing are depicted in Figures 2a and 2b, respectively. Across all data sets, the sociodemographic classification improves with more pretraining data. The learning curves in the classic probing do not flatten out, which indicates the potential for more research on the topic. We conclude that *with more pretraining data, more sociodemographic knowledge is present in the features.* This finding contrasts with other lower-level tasks, such as syntactic knowledge (Zhang et al., 2021b; Pérez-Mayos et al., 2021). As we hypothesized, though, it is similar to other higher-level language aspects, like common sense knowledge. Our control task, CoLA, exactly reflects this trend: the learning curve of



(a) Classic probing



(b) MDL probing

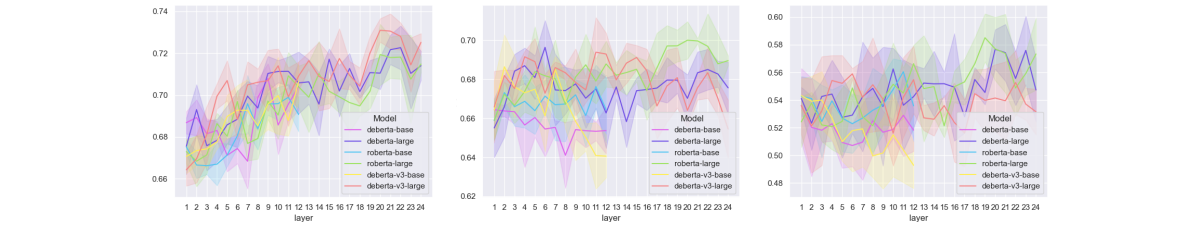
Figure 2: Classic and MDL probing results for RoBERTa models trained on varying amounts (1M–30B words) of pre-training data (RQ2).

predicting linguistic acceptability is much steeper.

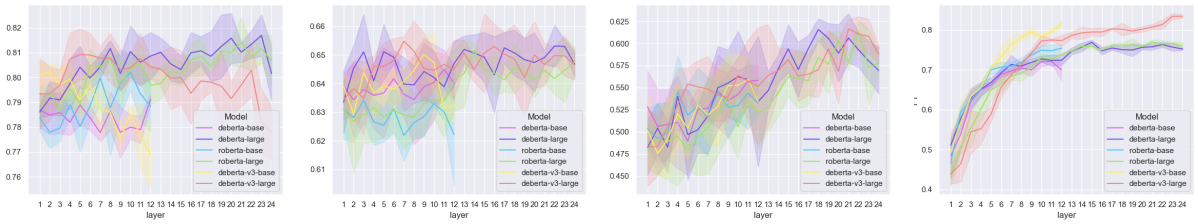
Cost-benefit Analysis. Inspired by Pérez-Mayos et al. (2021), we conduct a cost-benefit analysis. The authors base their estimate on the costs provided in Strubell et al. (2019). We follow their approach,[8] and approximate the financial costs of training a model with $60,948 / 30B words * #TrainingWords for each of the MiniBERTas, and the CO_2 emissions of each MiniBERTa model as 6,990 lbs / 30B * #TrainingWords. The final cost needs to be scaled with the number of pre-training procedures needed for model optimization reported by Warstadt et al. (2020) (10 times for the 1B MiniBERTa, 25 times for the other MiniBERTa models). In contrast to Pérez-Mayos et al. (2021), we also include RoBERTa *base* in our analysis and scale the costs accordingly. Table 2 shows the

[7] Publicly available at https://huggingface.co/nyu-mll/roberta-med-small-1M-1

[8] Note, that these are presumably a overestimates, as hardware has been getting cheaper and more power efficient.

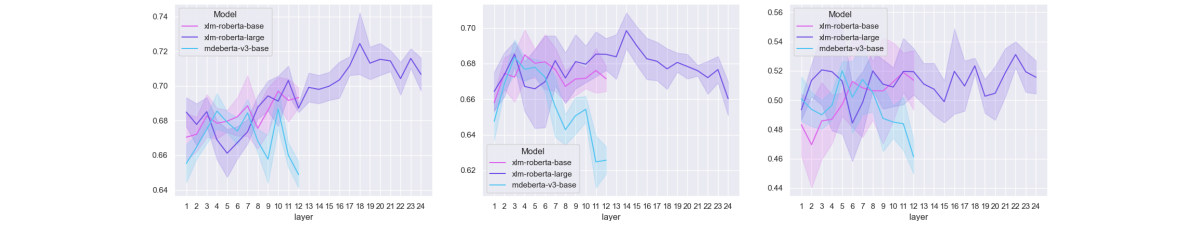


(a) Trustpilot (Age) (b) Trustpilot (Gender) (c) Facebook Congress

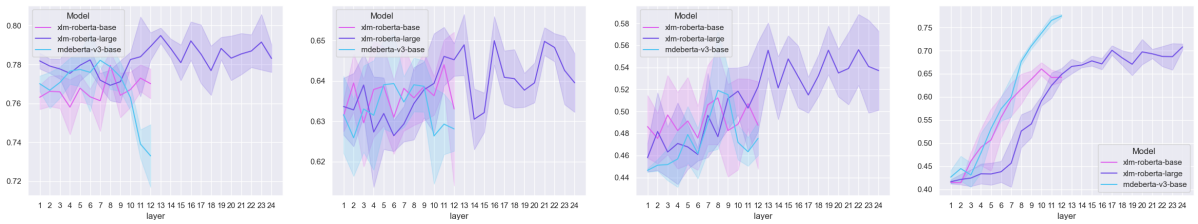


(d) Facebook Wiki (e) Fitocracy (f) Reddit (g) CoLA

Figure 3: Layer-wise F1-scores (average and standard deviation) for DeBERTa original and v3 *large* and *base* and RoBERTa *large* and *base* across 5 runs and 7 tasks ((a) Trustpilot Age to (g) CoLA).



(a) Trustpilot (Age) (b) Trustpilot (Gender) (c) Facebook Congress



(d) Facebook Wiki (e) Fitocracy (f) Reddit (g) CoLA

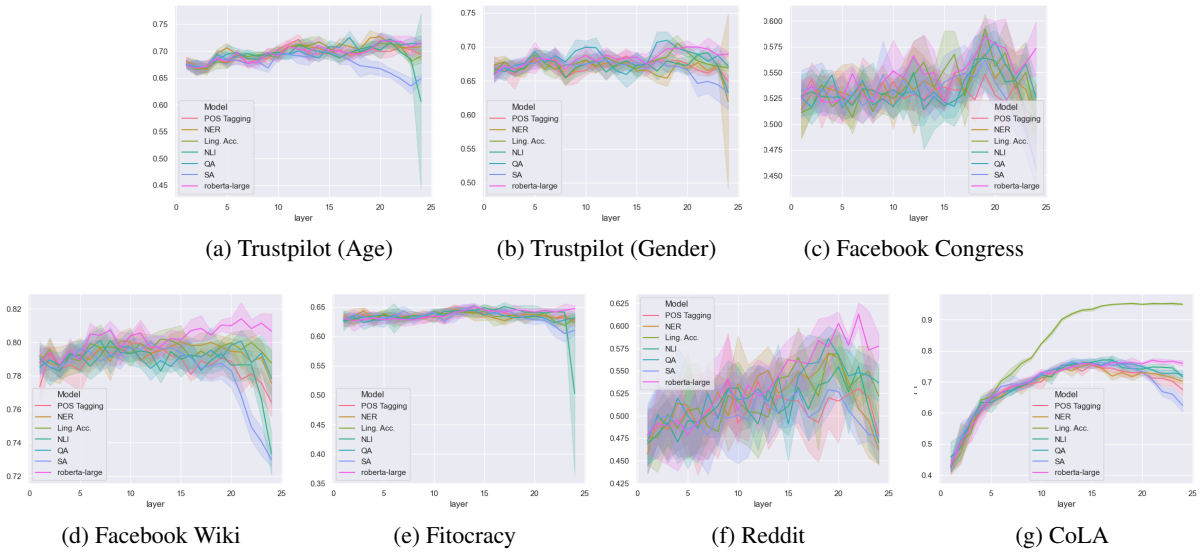
Figure 4: Results for our analysis of multilingual models (RQ4). We show F1-scores (average and standard deviation) across 5 runs on 7 tasks ((a) Trustpilot (Age) to (g) CoLA). The features we probe are extracted from different layers of XLM-RoBERTa *large*, XLM-RoBERTa *base*, and mDeBERTa *base*.

cost estimates and expected performance improvements. Between 1M and 1B tokens the expected gains flatten (see previous analysis), while the gains are lower than the ones reported by Pérez-Mayos et al. (2021) for syntactic tasks. However, with 30B we can expect a large performance improvement indicating that higher performance can only be expected at even higher financial and environmental costs. Given that the already high baseline costs, such a development is ethically problematic. Our results support the need for more research on sustainable NLP, especially when tasks require in-depth language understanding.

5.4 RQ3: Where is sociodemographic knowledge located?

We test embeddings extracted from different layers.

Approach. In the previous experiments, we followed the standard approach and pooled representations from the last layer of the Transformer. In contrast, here we test the average pooled representations from *each* layer $n \in [1 : \text{num\_layers}]$, where



(a) Trustpilot (Age) (b) Trustpilot (Gender) (c) Facebook Congress

(d) Facebook Wiki (e) Fitocracy (f) Reddit (g) CoLA

Figure 5: Results of our STILTs analysis (RQ5) in terms of F1 scores (average and standard deviation) for 5 runs across 7 tasks ((a) Trustpilot Age to (g) CoLA). The features we probe are extracted from the original RoBERTa *large*, and RoBERTa *large* models, which we obtain from fine-tuning RoBERTa on 6 tasks: Part-of-Speech Tagging (POS Tagging), Named Entity Recognition (NER), Linguistic Acceptability Prediction (Ling. Acc.), Natural Language Inference (NLI), Question Answering (QA), and Sentiment Analysis (SA).

num_layers corresponds to the number of layers in the model. We test RoBERTa and DeBERTa (original and v3) in the *large* and *base* configurations.

Results. We show the results in Figures 3a–3g. Note the relatively high standard deviations compared to the overall performance range. The exception to this observation is again CoLA, our control task (Figure 3g). The tendency seems to be that higher layers offer better representations for sociodemographic classification (e.g., Trustpilot (Age) in Figure 3a, Reddit (Gender) in Figure 3f), but performance improvement across layers is much more skewed for CoLA than for the sociodemographic probing tasks. Especially for DeBERTa v3 *base*, the probing results are often better for lower model layers (e.g., Figure 3c). This runs counter to Tenney et al. (2019), who showed higher-level semantic knowledge to be encoded in the higher layers of BERT. We conclude that *sociodemographic knowledge is much less localized* in PLMs than lower-level knowledge. This finding corresponds to the observation that different sociodemographic factors are expressed in different ways (Johannsen et al., 2015). As in our experiments for answering RQ1, DeBERTa *large* v3 has superior knowledge about lower-level linguistic aspects, but not sociodemographic knowledge.

5.5 RQ4: Does the sociodemographic knowledge in multilingual models differ?

Hung et al. (2022) recently showed that straightforward attempts to (socio)demographic adaptation of multilingual Transformers can lead to a better separation of representation areas according to input text languages and not according to author demographics. This leads us to question whether the multilingual signal significantly affects the encoding of sociodemographic knowledge in the models. We probe multilingual PLMs for their encoding of sociodemographics in English and further validate our previous findings.

Approach. We use multilingual versions of RoBERTa and DeBERTa available on Huggingface: XLM-RoBERTa in *large* and *base* configuration and mDeBERTa v3 in *base* configuration.[9]

Result. We only show the classic probing results (Figure 4, see Appendix for MDL). While the scores are slightly lower than for monolingual PLMs, they are generally in-line with our findings from RQ2: sociodemographic knowledge is less localized than that for CoLA. While for XLM-RoBERTa *large* and *base* higher layers encode the sociodemographics, the results of DeBERTa v3 *base* show an opposite trend. We conclude that *the localization of sociodemographic knowl-*

[9]No *large* configuration of mDeBERTa was available

edge in multilingual models follows their monolingual counterparts. The layer-wise behavior of DeBERTa v3 *base* is an additional pointer to the effect of the training regimes on the sociodemographic knowledge encoded in PLMs.

5.6 RQ5: What is the effect of STILTs on the encoding of the knowledge?

We explore the effect of STILTs on the sociodemographic knowledge encoded in the layers.

Approach. We use the encoders of readily fine-tuned RoBERTa *large* models from the Huggingface repository trained on the following tasks and data sets: POS tagging and dependency parsing on UPOS, named entity recognition, natural language inference on MNLI, question answering on SQuaD v.2., sentiment analysis on SST2, and linguistic acceptability prediction on our control task CoLA.

Results. See Figure 5 for the classic probing results (MDL probing results in the Appendix). Unsurprisingly, supplementary training on Linguistic Acceptability Prediction (=our control task CoLA), leads to superior representations for CoLA probing (Figure 5g). This effect is clearly visible from layer 5 onwards, indicating that the top 19 layers become specialized during the STILTs fine-tuning. In contrast, the selected STILTs tasks do not improve the sociodemographic knowledge in the representations (e.g., QA STILTs for gender prediction in Trustpilot) or even reduce that knowledge (e.g., NLI and SA STILTs for gender prediction in FB Wiki (Figure 5d)). The results suggest that sociodemographic knowledge is overwritten during STILTs. Interestingly, this effect mostly occurs on the last 5 to 10 layers (e.g., Trustpilot Age prediction from layer 10 (Figure 5a), and much more gently than the CoLA improvement.

6 Conclusion

Sociodemographic aspects shape our language and are thus important factors to model in language technology. However, despite a plethora of works probing PLMs for various types of knowledge, we know little about these higher-level aspects of language. We present SOCIOPROBE to understand *whether*, *when*, and *where* PLMs encode sociodemographic knowledge in their representations. We find that sociodemograaphic knowledge is located in PLMs, but much more diffuse than lower-level aspects. In the future, we will extend our analysis to languages other than English. We hope that our findings will fuel further research towards human-like language understanding.

Acknowledgements

The work of Anne Lauscher is funded under the Excellence Strategy of the German Federal Government and the Länder. This work is in part funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). At the time of writing, AL, FB, and DH were members of the Data and Marketing Insights unit of the Bocconi Institute for Data Science and Analysis

Limitations

Our work deals with predicting sociodemographic aspects from text, which should be considered sensitive information. Predictive methods can result in potentially harmful applications, e.g., in the context of user profiling. We acknowledge this potential for *dual use* (Jonas, 1984) of the data sets we use. However, in this work, we are interested in advancing NLP research towards a *better understanding of such fine-grained aspects of language and how they are already captured by our technology*. We believe that these insights will lead us toward fairer and more inclusive language technology. In contrast, we explicitly discourage the prediction of sensitive attributes from text for harmful purposes.

Further, we acknowledge that our work is limited in that the data sets available to us model gender as a binary variable, which does not reflect the wide variety of possible identites along the gender spectrum and beyond (Lauscher et al., 2022). However, we are not aware of other suitable data sets without this limitation. We have reason to believe, though, that even the findings derived from a binary view on gender (as well as for age) can provide an initial understanding of how language varies, and that any results will hold under a more sophisticated modeling of the problem.

An additional limitation of our work comes from the pre-trained models we used. All the models tested are easily-downloadable single-GPU models that have been pre-trained on general-purpose data. We acknowledge that results might differ for models that were of bigger capacity and pre-trained on data from other and more specific domains, e.g., social media. The same argument can be made about the architectures used. We mainly focused on BERT-like models trained via MLM, which are

only a subset of the language models proposed in the literature. We leave the exploration of these effects (e.g., pre-training objective) for future work.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR*, Toulon, France.

Meriem Beloucif and Chris Biemann. 2021. Probing pre-trained language models for semantic attributes and their values. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2554–2559, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. 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.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Steven Cao, Victor Sanh, and Alexander Rush. 2021. Low-complexity probing via finding subnetworks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–966, Online. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. 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.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and gender*. Cambridge University Press.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

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

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Chia-Chien Hung, Anne Lauscher, Dirk Hovy, Simone Paolo Ponzetto, and Goran Glavaš. 2022. Can demographic factors improve text classification? revisiting demographic adaptation in the age of transformers. *arXiv preprint arXiv:2210.07362*.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89, Minneapolis, USA. Association for Computational Linguistics.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China. Association for Computational Linguistics.

Hans Jonas. 1984. *The imperative of responsibility: In search of an ethics for the technological age*. University of Chicago Press. Original in German: Prinzip Verantwortung.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2021. Probing multilingual language models for discourse. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 8–19, Online. Association for Computational Linguistics.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*

(Volume 2: Short Papers), pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, 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. *arXiv preprint arXiv:1907.11692*.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does RoBERTa know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2017. Language-independent gender prediction on Twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 1–6, Vancouver, Canada. Association for Computational Linguistics.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

Matej Martinc and Senja Pollak. 2018. Reusable workflows for gender prediction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Dong Nguyen, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Onkar Pandit and Yufang Hou. 2021. Probing for bridging inference in transformer language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4153–4163, Online. Association for Computational Linguistics.

Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner. 2021. How much pretraining data do language models need to learn syntax? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1571–1582.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Tiago Pimentel and Ryan Cotterell. 2021. A Bayesian framework for information-theoretic probing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2869–2887, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Archiki Prasad and Preethi Jyothi. 2020. How accents confound: Probing for accent information in end-to-end speech recognition systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3739–3753, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Vinit Ravishankar, Lilja Øvrelid, and Erik Velldal. 2019. Probing multilingual sentence representations with X-probe. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 156–168, Florence, Italy. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 763–772, Portland, Oregon, USA. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

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

Naomi Shapiro, Amandalynne Paullada, and Shane Steinert-Threlkeld. 2021. A multilabel approach to morphosyntactic probing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4486–4524, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. 2020. Probing for referential information in language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4177–4189, Online. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Peter Trudgill. 2000. *Sociolinguistics: An introduction to language and society*. Penguin UK.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, Washington, USA. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

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.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Chien-Sheng Wu and Caiming Xiong. 2020. Probing task-oriented dialogue representation from language

models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5036–5051, Online. Association for Computational Linguistics.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. 2021a. Sociolectal analysis of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4588, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021b. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Models Used

We provide an overview on all models we have used in this study. They are available on the Huggingface Hub: https://huggingface.co.

A.1 Models Used for RQ1

For our base experiment we have used the following models.

- `roberta-base`: 12 layers, 12 attention heads, hidden size of 768
- `roberta-large`: 24 layers, 16 attention heads, hidden size of 1024
- `microsoft/deberta-v3-base`: 12 layers, 12 attention heads, hidden size of 768
- `microsoft/deberta-v3-large`: 24 layers, 16 attention heads, hidden size of 1024
- `microsoft/deberta-v3-xsmall`: 12 layers, 6 attention heads, hidden size of 384
- `microsoft/deberta-v3-small`: 6 layers, 12 attention heads, hidden size of 768
- `microsoft/deberta-base`: 12 layers, 12 attention heads, hidden size of 768
- `microsoft/deberta-large`: 24 layers, 16 attention heads, hidden size of 1024

A.2 Models Used for RQ2

For investigating the amount of pre-training data needed, we have used the suite of MiniBERTas, and RoBERTa *base*.

- `roberta-base`: 12 layers, 12 attention heads, hidden size of 768
- `nyu-mll/roberta-base-1B-1`: 12 layers, 12 attention heads, hidden size of 768
- `nyu-mll/roberta-base-1B-2`: 12 layers, 12 attention heads, hidden size of 768
- `nyu-mll/roberta-base-1B-3`: 12 layers, 12 attention heads, hidden size of 768
- `nyu-mll/roberta-base-100M-1`: 12 layers, 12 attention heads, hidden size of 768
- `nyu-mll/roberta-base-100M-2`: 12 layers, 12 attention heads, hidden size of 768
- `nyu-mll/roberta-base-100M-3`: 12 layers, 12 attention heads, hidden size of 768
- `nyu-mll/roberta-base-10M-1`: 12 layers, 12 attention heads, hidden size of 768
- `nyu-mll/roberta-base-10M-2`: 12 layers, 12 attention heads, hidden size of 768
- `nyu-mll/roberta-base-10M-3`: 12 layers, 12 attention heads, hidden size of 768
- `nyu-mll/roberta-med-small-1M-1`: 6 layers, 8 attention heads, hidden size of 512
- `nyu-mll/roberta-med-small-1M-2`: 6 layers, 8 attention heads, hidden size of 512
- `nyu-mll/roberta-med-small-1M-3`: 6 layers, 8 attention heads, hidden size of 512

A.3 Models Used for RQ3

We investigated the layer-wise knowledge of the following models.

- `roberta-base`: 12 layers, 12 attention heads, hidden size of 768
- `roberta-large`: 24 layers, 16 attention heads, hidden size of 1024
- `microsoft/deberta-v3-base`: 12 layers, 12 attention heads, hidden size of 768
- `microsoft/deberta-v3-large`: 24 layers, 16 attention heads, hidden size of 1024
- `microsoft/deberta-base`: 12 layers, 12 attention heads, hidden size of 768
- `microsoft/deberta-large`: 24 layers, 16 attention heads, hidden size of 1024

A.4 Models Used for RQ4

As multilingual counter-parts, we employed the following models.

- `xlm-roberta-base`: 12 layers, 12 attention heads, hidden size of 768
- `xlm-roberta-large`: 24 layers, 16 heads, hidden size of 1024
- `microsoft/mdeberta-v3-base`: 12 layers, 12 attention heads, hidden size of 768

A.5 Models Used for RQ5

Finally, we ran the STILT experiment with the following models.

- `roberta-large`: 24 layers, 16 heads, hidden size of 1024
- `KoichiYasuoka/roberta-large-english-upos`: 24 layers, 16 heads, hidden size of 1024
- `Jean-Baptiste/roberta-large-ner-english`: 24 layers, 16 heads, hidden size of 1024
- `cointegrated/roberta-large-cola-krishna2020`: 24 layers, 16 heads, hidden size of 1024
- `roberta-large-mnli`: 24 layers, 16 heads, hidden size of 1024
- `navteca/roberta-large-squad2`: 24 layers, 16 heads, hidden size of 1024
- `howey/roberta-large-sst2`: 24 layers, 16 heads, hidden size of 1024

B Additional Results

We provide the additional results for MDL probing.

B.1 Additional Results for RQ3

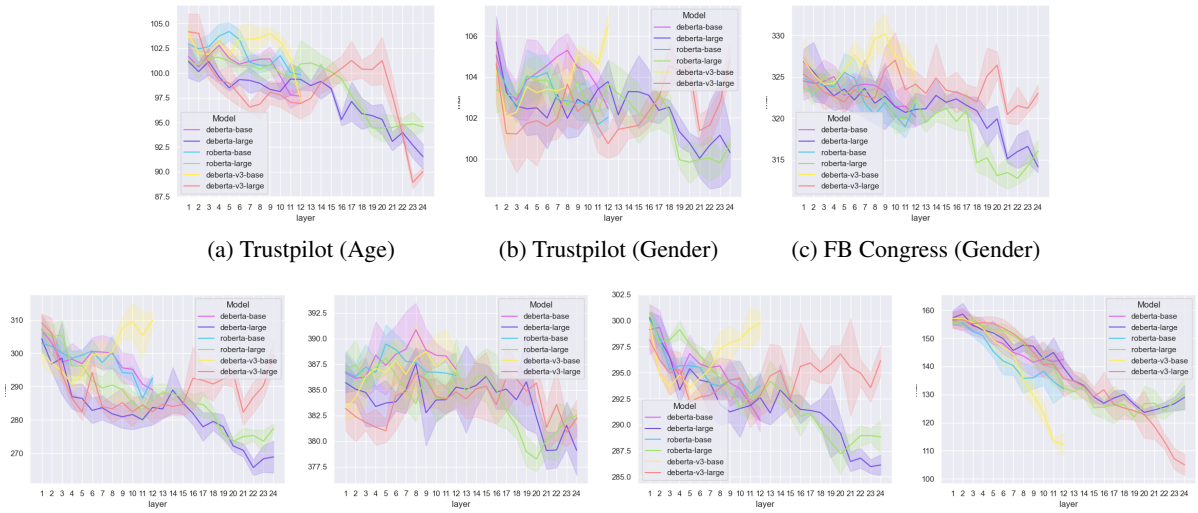
The layer-wise analysis for MDL probing is provided in Figure 6.

B.2 Additional Results for RQ4

We show the MDL results for the multilingual analysis in Figure 7.

B.3 Additional Results for RQ5

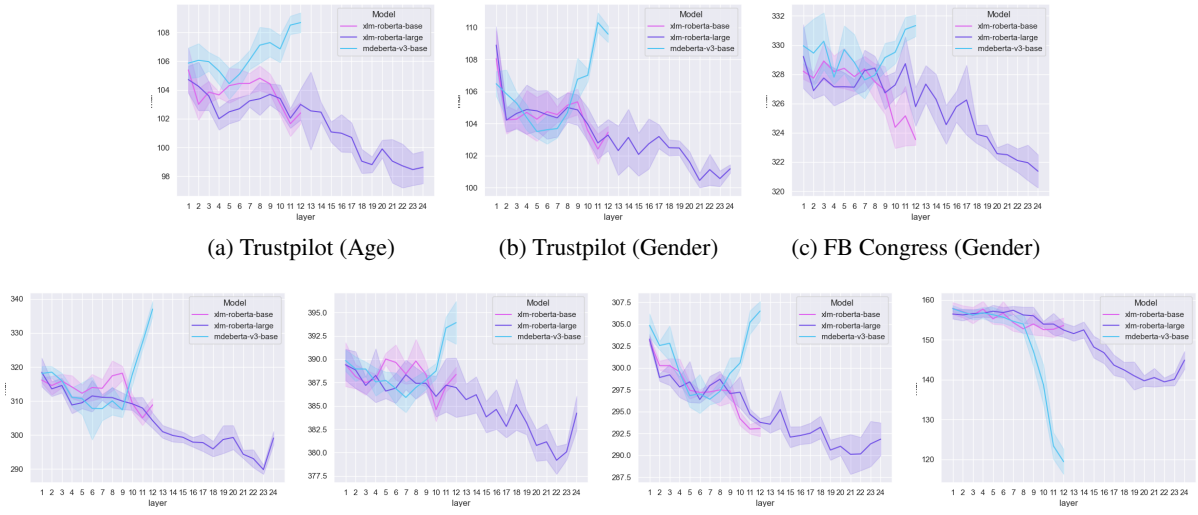
We provide the MDL probing results for our STILT analysis in Figure 8.



(a) Trustpilot (Age) (b) Trustpilot (Gender) (c) FB Congress (Gender)

(d) FB Public (Gender) (e) Fitocracy (Gender) (f) Reddit (Gender) (g) CoLA

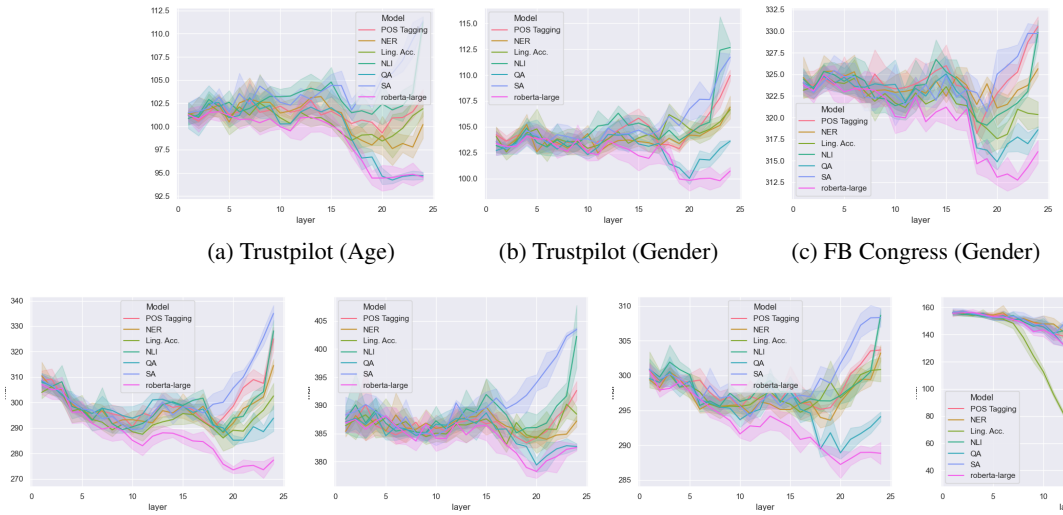
Figure 6: Results showing our layer-wise analysis of DeBERTa original and v3 *large* and *base* and RoBERTa *large* and *base* in terms of average and standard deviation of the MDL for 5 runs across 7 tasks ((a) Trustpilot Age to (g) CoLA)



(a) Trustpilot (Age) (b) Trustpilot (Gender) (c) FB Congress (Gender)

(d) FB Public (Gender) (e) Fitocracy (Gender) (f) Reddit (Gender) (g) CoLA

Figure 7: Results for our analysis of multilingual models in terms of average and standard deviation of the MDL scores for 5 runs across 7 tasks ((a) Trustpilot (Age) to (g) CoLA) for features extracted from different layers of XLM-RoBERTa *large* and *base* and mDeBERTa *base*.



(a) Trustpilot (Age) (b) Trustpilot (Gender) (c) FB Congress (Gender)

(d) FB Public (Gender) (e) Fitocracy (Gender) (f) Reddit (Gender) (g) CoLA

Figure 8: Results for our Supplementary Training on Intermediate Labeled Tasks (STILT) analysis. We show the probing results in terms of average and standard deviation of the MDL scores for 5 runs across 7 tasks ((a) Trustpilot Age to (g) CoLA) for features extracted from the original RoBERTa *large*, and RoBERTa *large* fine-tuned on 6 tasks: Part-of-Speech Tagging (POS Tagging), Named Entity Recognition (NER), Linguistic Acceptability Prediction (Ling. Acc.), Natural Language Inference (NLI), Question Answering (QA), and Sentiment Analysis (SA).