SocioProbe: What, When, and Where Language Models Learn about Sociodemographics
Anne Lauscher1, Federico Bianchi2, Samuel Bowman3, and Dirk Hovy4
1Data Science Group, University of Hamburg, Germany
2StanfordNLP, Stanford University, CA, USA
3New York University, NY, USA
4MilaNLP, 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., lexi-co-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
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 sociographics 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 extent 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 sociographics 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 subtractive (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 (Vušić 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.

---

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

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Textual Domain</th>
<th>Dimension</th>
<th>Label</th>
<th># Instances</th>
<th>% Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trustpilot</td>
<td>Product Reviews</td>
<td>Gender</td>
<td>Man</td>
<td>5349</td>
<td>49.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Woman</td>
<td>50.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>Young</td>
<td>5269</td>
<td>52.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Old</td>
<td>47.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTGender</td>
<td>Facebook Posts (Congress Members)</td>
<td>Gender</td>
<td>Man</td>
<td>510135</td>
<td>75.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Woman</td>
<td>24.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Facebook Posts (Public Figures)</td>
<td>Gender</td>
<td>Man</td>
<td>133,017</td>
<td>33.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Woman</td>
<td>66.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fitocracy Posts</td>
<td>Gender</td>
<td>Man</td>
<td>318,535</td>
<td>54.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Woman</td>
<td>45.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reddit Posts</td>
<td>Gender</td>
<td>Man</td>
<td>1,453,512</td>
<td>79.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Woman</td>
<td>20.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Trustpilot (Hovy, 2015). Trustpilot 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 dataset. 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, Facebook (posts from politicians and public figures), and Fitocracy. 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: \( \{(x_j, y_j)\}_{i=1}^{t} \) 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:

\[
\text{MDL} = t_1 \log_2 2 - \sum_{i=1}^{10} \mathcal{L}_{i+1},
\]

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

7904
5.2 RQ1: To what extent 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
<table>
<thead>
<tr>
<th># Tokens</th>
<th>Costs ($)</th>
<th>CO₂ (lbs)</th>
<th>$\mu$Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1M</td>
<td>50</td>
<td>5.825</td>
<td>–</td>
</tr>
<tr>
<td>10M</td>
<td>500</td>
<td>58.250</td>
<td>+2.61</td>
</tr>
<tr>
<td>100M</td>
<td>5,075</td>
<td>582.500</td>
<td>+1.98</td>
</tr>
<tr>
<td>1B</td>
<td>20,320</td>
<td>2,330.000</td>
<td>+0.30</td>
</tr>
<tr>
<td>30B</td>
<td>609,600</td>
<td>69,900.000</td>
<td>+8.56</td>
</tr>
</tbody>
</table>

Table 2: Results of our cost-benefit analysis. We show financial costs (Costs ($)) and CO₂ emissions (CO₂ (lbs)), gain is average F1-measure increase over the extent smaller model across all data sets and models ($\mu$Gain).

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 MiniBERT-Tas (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 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 CO₂ 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

---

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

⁸Note, that these are presumably a overestimates, as hardware has been getting cheaper and more power efficient.
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 $n \in [1 : \text{num} \_ \text{layers}]$, where
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.

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.\footnote{No large configuration of mDeBERTa was available}

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-

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.

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).
edge in multilingual models follows their mono-
lingual counterparts. The layer-wise behavior of
DeBERTa v3 base is an additional pointer to the
effect of the training regimes on the sociode-
graphic 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 sociode-
graphic knowledge encoded in the layers.

Approach. We use the encoders of readily fine-
tuned RoBERTa large models from the Hugging-
face 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 re-
sults (MDL probing results in the Appendix). Un-
surprisingly, supplementary training on Linguistic
Acceptability Prediction (=our control task CoLA),
leads to superior representations for CoLA prob-
ing (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 im-
prove the sociodemographic knowledge in the rep-
resentations (e.g., QA STILTs for gender predic-
tion 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 dur-
ing STILTs. Interestingly, this effect mostly occurs
on the last 5 to 10 layers (e.g., Trustpilot Age pre-
diction 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 lan-
guage. We present SOCIOPROBE to understand
whether, when, and where PLMs encode sociode-
mographic knowledge in their representations. We
find that sociodemographic 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 Govern-
ment 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 sen-
sitive 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 advanc-
ing 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 con-
trast, 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 spec-
trum 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 ini-
tial 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 mod-
els 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


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, 8 attention heads, hidden size of 512
- 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
- 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, 8 attention heads, hidden size of 512
- 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.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.
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).

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.
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).