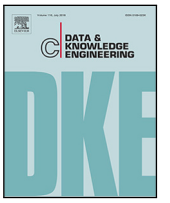
Data & Knowledge Engineering 148 (2023) 102235

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak



Towards comparable ratings: Exploring bias in German physician reviews

Joschka Kersting *, Falk Maoro, Michaela Geierhos

University of the Bundeswehr Munich, Research Institute CODE, Neubiberg, Germany

ARTICLE INFO

Keywords:
Language model fairness
Aspect phrase classification
Grade prediction
Physician reviews

ABSTRACT

In this study, we evaluate the impact of gender-biased data from German-language physician reviews on the fairness of fine-tuned language models. For two different downstream tasks, we use data reported to be gender biased and aggregate it with annotations. First, we propose a new approach to aspect-based sentiment analysis that allows identifying, extracting, and classifying implicit and explicit aspect phrases and their polarity within a single model. The second task we present is grade prediction, where we predict the overall grade of a review on the basis of the review text. For both tasks, we train numerous transformer models and evaluate their performance. The aggregation of sensitive attributes, such as a physician's gender and migration background, with individual text reviews allows us to measure the performance of the models with respect to these sensitive groups. These group-wise performance measures act as extrinsic bias measures for our downstream tasks. In addition, we translate several gender-specific templates of the intrinsic bias metrics into the German language and evaluate our fine-tuned models. Based on this set of tasks, fine-tuned models, and intrinsic and extrinsic bias measures, we perform correlation analyses between intrinsic and extrinsic bias measures. In terms of sensitive groups and effect sizes, our bias measure results show different directions. Furthermore, correlations between measures of intrinsic and extrinsic bias can be observed in different directions. This leads us to conclude that gender-biased data does not inherently lead to biased models. Other variables, such as template dependency for intrinsic measures and label distribution in the data, must be taken into account as they strongly influence the metric results. Therefore, we suggest that metrics and templates should be chosen according to the given task and the biases to be assessed.

1. Introduction

Feedback is essential for evaluating and comparing products and services offered. It helps businesses understand the interests and opinions of their customers. More often than not, feedback is unstructured and therefore difficult to parse. To analyze feedback data, large language models (LLMs) can be used to extract explicit information from unstructured text. This information can then be used in, for example, recommender systems. Aspect-based sentiment analysis (ABSA) is one way to structure unstructured data [1]. However, the use of LLMs in pipelines carries the risk of bias, such as underrepresented gender, ethnic, or religious groups. This is particularly important when reviewing people. Therefore, it is necessary to identify potential biases and find ways to avoid disadvantaging or discriminating.

* Corresponding author.
E-mail addresses: joschka.kersting@unibw.de (J. Kersting), falk.maoro@unibw.de (F. Maoro), michaela.geierhos@unibw.de (M. Geierhos).

https://doi.org/10.1016/j.datak.2023.102235
Received 30 November 2022; Received in revised form 16 September 2023; Accepted 9 October 2023
Available online 11 October 2023

Kauff et al. [2] investigated whether ethnic and gender bias could be found in the ratings of male and female physicians with a migration background. Their study suggests that general practitioners with non-German names were rated lower than general practitioners with German names. Their analyses also showed that female physicians were rated less positively. The question is whether these data biases have an effect on fine-tuned LLMs since these ethnic and gender biases have been demonstrated [2]. Therefore, we will investigate the effects of fine-tuning German or multilingual LLMs on the biased data in several downstream tasks.

In order to investigate the impact of gender-biased data on language models, we perform a number of fairness measurements on pre-trained and fine-tuned language models. Our study is based on data that is reported to be gender-biased and includes annotations regarding the migration background of physicians. We focus on assessing gender bias and evaluating migration bias where applicable, as some metrics or their respective templates are not suitable for testing migration bias. We aggregate the data with annotations for different purposes and fine-tune the models on two tasks (i.e., ABSA and grade prediction). In particular, we will investigate whether language models are able to learn biases from the underlying data and how these biases affect the model's performance. For the regression (e.g., grade prediction) and token classification tasks (e.g., ABSA), we also investigate whether the results of intrinsic bias measures correlate with extrinsic bias measures. In addition, we present a new condensed approach to the modeling of ABSA.

Thus, this study is organized as follows: After introducing the state of the art in online physician data, language models, measuring bias in language models, and ABSA in Section 2, we present the data we use in Section 3. The data has been enriched in several publications and comes from a raw corpus of physician review websites. Then, in Section 4, we present two tasks on which we train the language models. The first task is a condensed version of the ABSA approach, and the second task is the prediction of the grades that are given for text reviews. Both tasks are evaluated using appropriate measures. In Section 5, we perform bias analyses on all pre-trained and fine-tuned models, first using intrinsic bias metrics and second using extrinsic bias metrics. Finally, in Section 6, we discuss our results and explore the correlations among the fairness measures.

2. Related literature

This work covers a variety of data, tasks, language models, and bias evaluation techniques. We present the state of the art relevant to the study, starting with physician review data in Section 2.1. We then present language models being used in this study in Section 2.2. Afterward, we discuss language model bias in Section 2.3, and finally, we review previous work on aspect-based sentiment analysis in Section 2.4.

2.1. Online physician reviews

Physician review websites (PRWs) are online platforms that allow patients to obtain objective and subjective information in the form of factual knowledge about physicians and other medical providers [3], as well as patient reviews consisting of text and quantitative ratings (such as grades). Therefore, PRWs are an important tool for patients to choose a suitable physician [4]. Physicians' ratings of PRWs are predominantly positive [5]. This applies to the quantitative evaluations as well as to the texts [6]. For some of the PRWs, it has been shown that the data have a migration and gender bias with respect to practitioners, resulting in less favorable ratings for women and for practitioners with non-German names [2]. The question is whether these data biases affect fine-tuned large language models (LLMs), since these ethical and gender biases have been demonstrated. Therefore, we will investigate the effects of fine-tuning German or multilingual LLMs on the biased data in several downstream tasks. For this reason, we focus our research on German-language PRWs.

2.2. Language models

Language models have a long history. A breakthrough came with word embedding models such as word2vec [7], GloVe [8], or FastText [9], which were trained to represent words as static vectors, called word embeddings. When embedding models are applied to sentences or text, the resulting vectors can be used to solve various language tasks, such as classification, regression, and topic modeling, or broader language understanding or generation tasks involving question answering or summarization. Based on the context of a word, more recent approaches to embedding models compute dynamic representations of words or tokens (subwords). These context-sensitive models are typically based on the transformer architecture [10], which has been reused and extended in various models combined with different datasets and training approaches. Popular models include the Universal Sentence Encoder [11], BERT [12], RoBERTa [13], XLM-RoBERTa [14], ALBERT [15], or ELECTRA [16].

A major advantage of such models is the ability to do transfer learning, where the training of a model is a two-stage process. In the first stage, pre-training, the model is trained on a self-supervised task, such as masked language modeling (MLM). In order to learn the syntax and semantics of the language, the model reconstructs randomly masked tokens within sentences. The pre-training phase is computationally expensive and data-intensive, as it involves the use of large corpora of unlabeled data. In the next step, fine-tuning, the model is adapted to solve a downstream task. The last layer of the model is replaced by a task-specific layer, called the model head. It can solve tasks such as classification, regression, or token classification. Another smaller dataset containing labels for that task is then used to train the model with the new layer. This technique of using a pre-trained model is efficient because the model does not need to learn language understanding when training a model on a task, only the task itself, and the pre-trained model and its weights can be reused and fine-tuned for multiple downstream tasks.

2.3. Language model bias

The bias in language models can be defined as two different types: the intrinsic bias and the extrinsic bias [17]. Since the transformer architecture [10] is widely used for transfer learning, where a model is first pre-trained on one dataset and then fine-tuned on a different task and dataset, there are two stages where bias can be introduced: pre-training and fine-tuning. In pre-training, the model is typically trained on a self-supervised task of language reconstruction. This pre-training introduces an intrinsic bias. It affects the implicit associations within the embedding representations generated by a model. If a model is trained on unbalanced data, this unbalance can be learned by the model and can negatively affect sensitive groups at inference time. Extrinsic bias is introduced at the fine-tuning stage, where the labeled data for the downstream task may also be biased toward sensitive groups. There are several studies showing gender [18–23], race [20,21,23], or age [23] bias. Appropriate metrics can measure both intrinsic and extrinsic bias. However, it has been shown, that not all metrics correlate with each other, which complicates the interpretation of bias [17,24].

2.3.1. Intrinsic bias metrics

Measuring intrinsic bias typically works by evaluating the associations a model generates toward sensitive groups. By computing the vector differences between word embeddings, relationships between the corresponding words can be represented. This shows that word embedding models learn gender relations. For example, if the vector difference of the words 'king' and 'man' is added to the word 'woman', the resulting embedding corresponds to the word 'queen' [25]. Furthermore, the word embeddings contain bias subspaces. By projecting occupations onto the *he-she* (gender) axis, computed by the vector difference of their embeddings, it was shown that gender bias was prevalent in both GLoVe [8] and Word2vec [7] models. Given two gender words and computing their vector difference, further analogies can be found by computing the cosine similarity of this bias axis to the vector difference of two sampled word embeddings, thus finding words that represent a bias in the given bias direction [26]. The calculation of vector differences with cosine similarity in word embeddings has also been used in the Word Embedding Association Test (WEAT), which measures the associations between two sets of target words and two sets of attribute words. By computing the differences in cosine similarity of all target words with the attribute words, one can observe whether a model associates one set of target words more strongly with one set of attributes than the other set of target words, implying a bias [27]. Word embedding models generate static embeddings for words, but do not take into account the context in which a word occurs. Context-aware language models, such as ELMo [28], BERT [12], and RoBERTa [13] generate different embeddings for the same word in different contexts. Since an embedding may be biased in one context and produce opposite results in another, simple word sets such as those used in WEAT are not sufficient for measuring bias in contextual word embeddings. The Sentence Encoder Association Test (SEAT) [20] extends the work of WEAT by using sentence templates, that are filled with the word sets. Three sets of tests have been published with SEAT. The Caliskan tests are derived from the WEAT publication [27] by filling the word sets into templates, that contain as few bias-inducing words as possible. The tests for the *Angry Black Woman* stereotype include tests that measure whether language models inherit the bias of associating more loud, angry, and imposing traits with black women compared to white women. The third set of tests includes double binds based on the stereotype, that successful women are less liked than equally successful men, especially when that success is in a male-oriented field [29]. Another metric that extends the work of WEAT is the Contextualized Embedding Association Test (CEAT), which uses the WEAT word sets to filter Reddit data for sentences containing those words. This filtering produces a large set of sentences that have a variety of contexts compared to the hand-crafted templates of SEAT [30].

Models trained on a Masked Language Modeling (MLM) target can be evaluated with other MLM-based metrics that also measure a model's associations. CrowS-Pairs (Crowdsourced Stereotype Pairs) is an English benchmark dataset of over 1500 sentence pairs across nine bias types, such as race, religion, and age. Each example contains two sentences, that are identical, except for the bias attribute, which is either a stereotype or an anti-stereotype. By masking all non-different tokens, one at a time, and summing the log-likelihoods, two sentences can be compared to see which sentence is most likely to be generated. In this way, the metric measures the percentage of examples where the model produces a higher likelihood for a stereotypical sentence [23]. The Context Association Test (CAT) benchmark takes a similar approach. This benchmark comes with StereoSet, a crowdsourced English dataset containing stereotypes for gender, profession, race, and religion. The intra-sentence examples are sentence triplets of nearly identical sentences that differ in one stereotypical, one anti-stereotypical, and one meaningless word. The inter-sentence examples contain a context sentence and three attribute sentences that are also either stereotypical, anti-stereotypical, or meaningless. The CAT measures whether a model preferentially generates stereotypical or anti-stereotypical sentences, resulting in the least bias when both probabilities are at 50%. It also takes into account the number of times a model would generate meaningless associations, in order to measure the language modeling ability [31].

Another template-based method for measuring bias in language models is DisCo (Discovery of Correlations), which also tests models on an MLM task. A total of 14 short English sentence templates are filled with either names or gendered nouns. The model predicts the masked token. If the top 3 generated tokens differ between sets, a bias is indicated [22]. The Log Probability Bias Score (LPBS) is a template-based method similar to DisCo. It consists of two templates filled with gender pronouns as targets and either job titles, positive or negative traits, or skills from three different datasets. The score measures the difference in the likelihood of generating a female or male sentence, while also correcting for the target probabilities [21]. The Bias Evaluation Corpus with Professions (BEC-Pro) is a dataset that extends the LPBS with English and German sentences. The results show that the method is not appropriate for German because gender is not only marked in pronouns but also in the corresponding occupational attributes. Therefore, it is not possible to make a comparison [19] because the attribute word, which already encodes gender information, influences the probability of the target word.

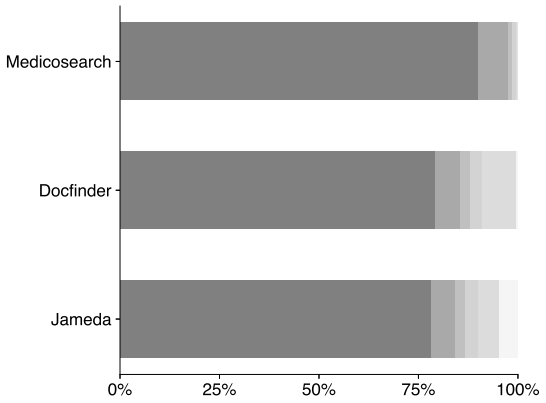


Fig. 1. Relative distribution of physician grades and stars in PRWs (darker = better, lighter = worse) [45].

2.3.2. Extrinsic bias metrics

Extrinsic bias is learned by training a model on a downstream task. Appropriate metrics are mostly task-specific due to the variety of downstream tasks that can be solved by embedding models. Measuring extrinsic bias involves relating performance metrics to sensitive attributes, such as gender or race, which extends common performance metrics such as precision, recall, loss, and so on. The prerequisite for this is that labels for these sensitive attributes are available for each prediction example.

Another problem is the large number of published fairness metrics, which must be carefully selected and interpreted, are sometimes impossible to combine, are mostly appropriate for binary classification tasks, and are inconsistently named. A categorization of these metrics includes statistical parity, disparate impact, equality of opportunity, calibration, and counterfactual fairness. In regression tasks where a loss function is minimized to produce the smallest difference between actual and predicted values, fairness can be achieved by using parity-based metrics that penalize a model for not treating sensitive attributes equally [32].

There are other extrinsic bias metrics, that focus on coreference resolution [33–35], classification [36], named entity recognition [37], translation [38], text generation [39,40], or information retrieval [18].

2.4. Aspect-based sentiment analysis

In unstructured text from PRWs, ABSA can be used to identify and distinguish opinions about rated physicians and their services. In preliminary research, we looked at implicit mentions of aspects [1,41–45]. By annotating a subset of reviews from German-language PRWs, a pipeline was implemented using several fine-tuned LLMs [46].

ABSA can be divided into three subtasks: Aspect Term Extraction (ATE), Aspect Category Classification (ACC), and Aspect Polarity Classification (APC) [47]. The goal of ATE is to extract all relevant aspect terms, that is, opinions, aspects, and other relevant words, from the text. While ACC classifies the ATE results into pre-defined aspect classes, APC deals with the question of whether an opinion is positive or negative (or neutral) with respect to an aspect. We will use the same PRW datasets as before [46], which are described in more detail in Section 3, for better comparability with our previous work and to investigate whether the trained models are prone to bias.

3. Datasets

The data used in this study is the raw data collected from PRWs. We segregated the data into subsets. Review texts, quantitative scores, and other physician information are included in the raw data presented in Section 3.1. The first subset includes only one of the three PRWs. It has been enriched with data on sensitive attributes, such as gender or migration background [2]. We describe this subset in Section 3.2. In Section 3.3 we present the second subset of data, that is annotated for ABSA. The examples are therefore annotated with hand-crafted annotations that cover aspect terms, aspect categories, and aspect polarities.

3.1. Raw dataset: PRW collection

The raw data was crawled from three PRWs (`Jameda.de`, `Docfinder.at`, and `Medicosearch.ch`), each from a German-speaking country (Germany, Austria, and Switzerland), between March and July 2018 [48]. It has already been shown that there are more than 400,000 physician profiles and more than 2,000,000 reviews across all of the platforms. `Jameda.de` contains the majority of profiles and reviews overall, with male physicians being rated slightly more often than female physicians [46]. The ratings given for the different categories are either in the form of German school grades (i.e., ranging from 1 to 6, with 1 being the best) or in the form of stars. The quantitative scores (i.e., grades or stars) used in the PRWs are unevenly distributed. Fig. 1 shows this distribution. For all three PRWs, it shows, how often physicians received the best grade, the second best grade, and so

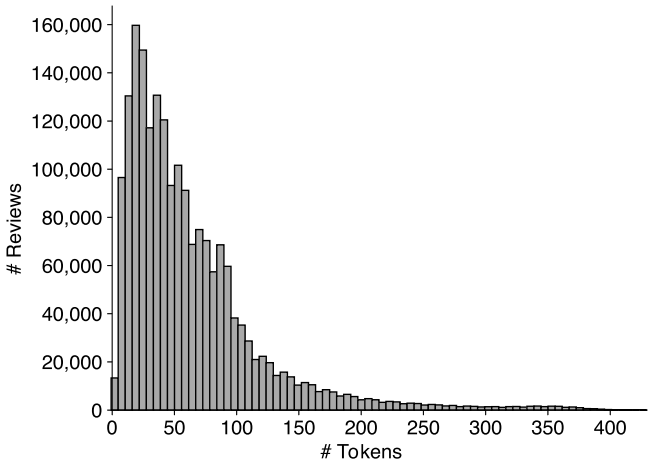


Fig. 2. Number of tokens per `Jameda.de` review [45].

Table 1
Number of physicians by gender and (non)immigrant background in SPAC.

	Male	Female	Sum
Migration background	9,460	7,224	16,684
No migration background	79,795	50,660	130,455
Sum	89,255	57,884	147,139

on. Most ratings are positive, as can be seen. Patients' use of the PRWs is not just an expression of negative feelings. However, the text ratings may contain insights not conveyed by the rating system because most of the ratings are positive.

Since the number of tokens can give an indication of the number of insights that can be extracted from the review text, we also examine the length of the reviews. Fig. 2 therefore shows the number of tokens per review. We combined the title and review text and used words instead of subwords. Fig. 2 shows that most reviews have less than 100, mostly less than 50 tokens, and are therefore rather short. As a result, patients may rate only a few aspects, or even just one aspect, or give a general recommendation. Furthermore, there are many reviews with more words that could describe other topics in detail.

3.2. Sensitive physician attributes corpus

Kauff et al. [2] have extended the patient review data to find relationships between physician ratings and physician ethnicity and gender. Several hypotheses were tested. These included that physicians with a migration background would receive worse ratings than German physicians without a migration background and that female physicians would receive worse ratings than male physicians [2]. We refer to this extended subset of our raw data as the "Sensitive Physician Attributes Corpus" (SPAC). For this dataset, a subset of the raw data was used and filtered for physicians only from the Jameda.de website. The data includes additional variables for 155,945 physicians. The variables are migration background, gender, zip code, and premium status on the site. Due to missing values for gender and migration background, we only use data for physicians with available values, resulting in 147,139 values. Neither gender nor migration background is evenly distributed across physicians, as can be seen in Table 1.

3.3. ABSA corpus

To create a dataset for ABSA based on supervised machine learning, a subset of the underlying raw data was annotated. We will refer to it as ABSAC. One approach was to use a pipeline of ATE, ACC, and APC. A single model trained on token classification solved ATE and ACC in combination. APC was a sequence classification problem. Thus, the annotated data contains annotations for classifying tokens and for classifying aspect polarity.

Sentences from the text reviews of the raw data were manually annotated for aspect phrases [1]. Several aspect classes were defined for this process, derived from the PRW's rating categories, resulting in 18 different classes [43,44]. The number of categories can be divided into groups dealing with different aspects of evaluation. Aspect classes that provide information about the physician are *friendliness*, *competence*, *time taken*, *explanation*, *alternative healing methods*, *treatment*, *care/commitment*, *overall/recommendation*, *child-friendliness*, and *relationship of trust*. Aspect classes that apply to practice staff are *care/commitment_staff*, *friendliness_staff*, *competence_staff*, and *accessibility by telephone*. The remaining categories cannot be applied to individuals. They concern *waiting time for an appointment*, *waiting time in the practice*, the *equipment* available in the practice, and the general *well-being* of the patients.

The annotations for ATE and ACC are spans of aspect terms, each with an aspect category. A subset of all ATE and ACC annotated sentences was further annotated for APC. By labeling the aspect terms as *positive* or *negative*, the polarity of the aspect terms was used to train APC models. Some examples of annotated data are shown in Section 4.1.

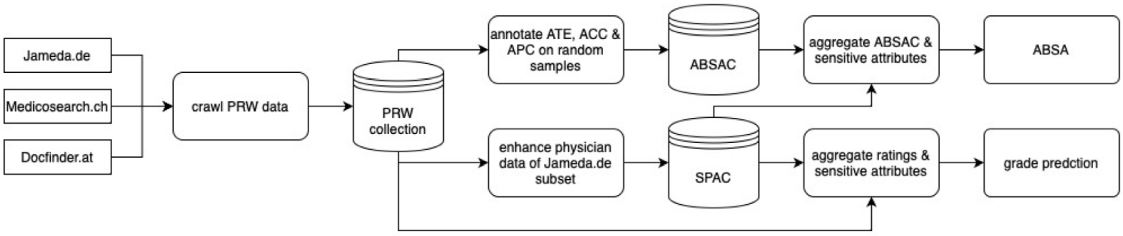


Fig. 3. Data processing pipeline for ABSA and grade prediction.

4. Classification and regression tasks

To investigate whether the uneven distribution of sensitive attributes affects the fairness of fine-tuned LLMs, we train classification and regression models on different downstream tasks described below. The fine-tuned models are later evaluated on several intrinsic and extrinsic bias metrics. Following our previous work, we present a condensed approach to ABSA as a classification task in Section 4.1. We also define a regression task, grade prediction based on physician review texts, in Section 4.2. The pipeline we use to process our data is visualized in Fig. 3. The pipeline is based on the PRW collection presented in Section 3.1. This data was annotated in previous work, resulting in the ABSAC and SPAC discussed in Section 3.3 and Section 3.2, respectively. By aggregating subsets of these datasets, we prepare the data used for the two tasks, ABSA and grade prediction.

4.1. Classification: ABSA

A three-task pipeline incorporating two separate fine-tuned LLMs was developed in previous work on the annotated data. ATE and ACC are solved by the first token classification model. The second model then classifies the polarity of each extracted aspect term or aspect category in APC [46]. There are a number of shortcomings with this approach, even though the labeling process may be simpler. First, the model receives the term and the sentence containing it, but no information about the aspect category, when predicting the polarity of an aspect. In addition, the APC model is also dependent on the token classifier. Thus, the polarity cannot be predicted if the token classifier does not recognize an aspect term. Furthermore, it is more complex to build such a pipeline, and it is more computationally intensive to train multiple models. Finally, one aspect that is relevant to this study, the bias, could be learned by both models. The identification of biases in two models that are built on top of each other could be difficult and has not been done so far.

Based on these deficiencies, we develop a combined approach for ABSA. For the three tasks of ATE, ACC, and APC, we use only the token classification model. By adding the polarity to each of the aspect categories, we are able to double the number of possible aspect categories. Thus, models can be trained to predict whether a token belongs to an aspect term, which category an aspect term belongs to, and its polarity. For that purpose, we use the annotated data presented in Section 3.3. One problem is the large number of aspect categories, which is even larger in combination with aspect polarities. To limit the complexity of the token classification task, we reduce the number of possible categories by dividing the annotated data into four different datasets. Each dataset deals with different evaluation aspects. The following list briefly describes the datasets and the categories they contain. Each category is available as *positive* and *negative*.

- **Dataset A** contains aspect targets that evaluate the physician. It summarizes relevant aspect classes dealing with physician behavior and competence [1,41,42]. Therefore, the categories *friendliness*, *competence*, *time taken*, and *explanation* have been labeled.
- In **Dataset B** six aspect classes are defined, which are also directed to the physician. The classes are *alternative healing methods*, *treatment*, *care/commitment*, *overall/recommendation*, *child-friendliness*, and *relationship of trust* [41].
- **Dataset C** contains all relevant aspect classes identified in the PRW data except those related to physicians. There are comments related to practice staff. The categories *care/commitment*, *friendliness*, *competence*, and *accessibility by telephone* were labeled [43].
- **Dataset D** contains annotated sentences related to waiting time, equipment, or a patient's subjective feelings about a doctor. The categories defined are *waiting time for an appointment*, *waiting time in the practice*, *equipment*, and *well-being* [44].

We aggregate the annotated data for ATE, ACC, and APC across all datasets. One difficulty is the availability of polarity annotations, since not all examples were annotated with polarities. In addition, for the purpose of bias exploration in Section 5, we aggregate the sentence data with the additional features of SPAC described in Section 3.2. Since the annotated sentences are derived from the text reviews of the raw data corpus of the three PRWs, and the SPAC data only contains annotations for physicians of the PRW *Jameda.de*, this further limits the number of examples. Table 2 shows the number of examples for each annotated dataset before and after aggregation with the polarity and the SPAC annotations.

In Fig. 4, we compare examples with and without aggregated polarity annotations. The aspect terms are highlighted in colored boxes, with the corresponding aspect category in bold. The examples show review sentences, each of which contains at least two

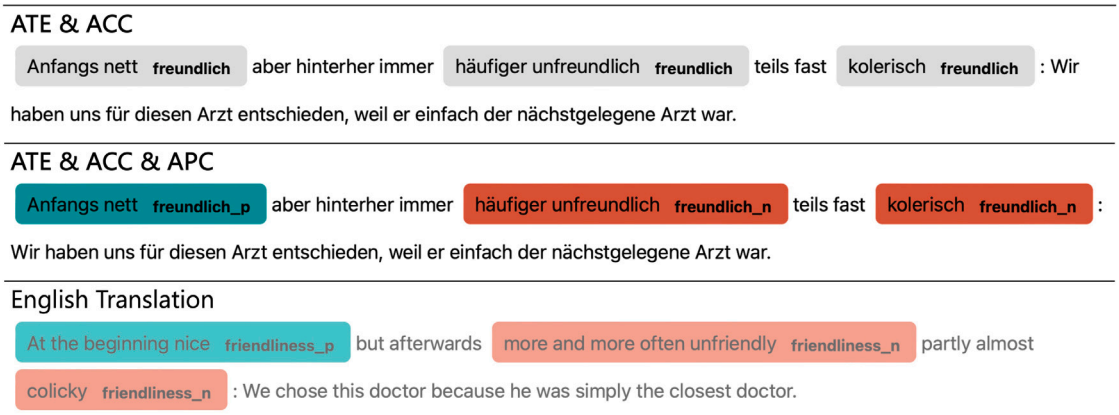


Fig. 4. Comparison of a sentence with annotations for ATE & ACC and APC.

Table 2
Number of annotated examples in the datasets combined with polarities and sensitive features.

Dataset	ATE & ACC	APC	APC with Gender & Background
A	6,334	1,939	1,409
B	6,597	1,878	1,265
C	7,996	1,951	1,350
D	7,414	1,781	1,177

aspect terms that share the same aspect category, but are different in terms of their polarity. Therefore, in our solution, the model predicts the polarity and the aspect class in one step.

Model training experiments are performed after processing the annotated data and aggregating the polarities. In previous work, XLM-RoBERTa, a multilingual transformer [14], was domain-trained on the PRW data. It performed best on the ATE & ACC tasks [43,44]. We also evaluate XLM-RoBERTa, although we do not perform any domain training before the fine-tuning step, for this combined task. We shuffle the aggregated data and take 80% as training data and 20% as test data. For phrase extraction and classification in one step, we use IO (Inside, Outside) tags for ATE and ACC [1,41–44]. We expect to benefit from learning the beginning and ending of sentences and their aspect classes and polarities. Technically, each word gets a tag in the form of "*I-friendliness_p*", "*I-friendliness_n*", or "*O*" for words outside of aspect phrases.

We show the results of XLM-RoBERTa-base and XLM-RoBERTa-large for each dataset in Table 3. The precision, recall and f1 metrics that we report are the macro-averages. As expected, the large model performs better than the base model for each measure. Even though the accuracy of each dataset ranges from 0.86 to 0.89, the F1 score, recall, and precision metrics show weak results. There are several reasons. First, as shown in Table 2, the training and test data are drastically reduced. Second, not all polarity-extended aspect classes are evenly distributed. This in turn results in categories that are never classified in the model evaluation, resulting in precision or recall values of 0. Since the macro-average scores are not weighted for uneven label distributions, the scores are quite low. Table 4 shows the distribution of the labels in the test set for dataset A. The O-tag indicates a token that is outside of a phrase. All other labels exist as positive and negative. It can be seen that the positive aspect categories are more frequent than the negative aspect categories. Moreover, the *O-tags* corresponding to non-aspect terms have the highest proportion of all labels. This pattern of uneven distribution is observed in all datasets, as shown in the appendix. Furthermore, the group-wise distribution of tags is also uneven. There are far more tokens for male physicians than for female physicians, and far more tokens for native-born physicians than for immigrated physicians. This goes to the point that there are no tokens for the tags *I-explanations_n* and *I-time_taken_n*.

4.2. Regression: Grade prediction

In addition to the overall ratings of physicians, the PRW collection also contains all reviews, including the text and different rating categories. Therefore, we can map the gender and migration background features from SPAC, added to the majority of physicians on the *Jameda.da* platform [2], to each physician's review. This aggregation resulted in 1,402,684 reviews, which we used in an 80:20 split for training and evaluation. To investigate whether a physician's gender or migration background influences a model's prediction, we run experiments to predict the overall grade for a text review with a fine-tuned language model. This allows a thorough analysis of the correlations between the predicted grades and the sensitive attributes in Section 5. The grades on the *Jameda.de* platform range from 1 to 6. For modeling purposes, we scaled the grades to a range from −1 to 1 to match the prediction

Table 3
Results for the combined ABSA approach.

Dataset	Model	Accuracy	Precision	Recall	F1
A	xlm-roberta-base	0.83	0.45	0.34	0.35
	xlm-roberta-large	**0.88**	**0.72**	**0.57**	**0.60**
B	xlm-roberta-base	0.81	0.41	0.24	0.26
	xlm-roberta-large	**0.86**	**0.50**	**0.44**	**0.45**
C	xlm-roberta-base	0.84	0.34	0.27	0.26
	xlm-roberta-large	**0.88**	**0.60**	**0.55**	**0.57**
D	xlm-roberta-base	0.84	0.39	0.38	0.38
	xlm-roberta-large	**0.89**	**0.67**	**0.71**	**0.68**

Table 4
Number of tokens per tag in the gender-annotated test subset of dataset A.

Tag	Polarity	Male	Female	Native	Migrant	Overall	Percentage
O		3,828	1,594	4,541	881	5,422	77.0%
Competence	P	367	167	439	95	534	7.6%
	N	17	32	44	5	49	0.7%
Explanation	P	152	41	151	42	193	2.7%
	N	9	23	32	0	32	0.5%
Friendliness	P	244	118	306	56	362	5.1%
	N	112	42	142	12	154	2.2%
Time Taken	P	165	91	207	49	256	3.6%
	N	28	8	36	0	36	0.5%

Table 5
Results for fine-tuned BERT grade prediction models.

Model	MSE	RMSE
deepset/gbert-base	0.070	0.265
deepset/gbert-large	**0.066**	**0.257**

range of the model. Because of this predictable range, the task is a constrained regression problem. We fine-tuned pre-trained BERT-based models by using the pooled embeddings and adding a tanh layer so that the models produce outputs between −1 and 1. The model is given the full review text, as opposed to ABSA where only a single sentence is used. The primary metric used to measure the model performance is the mean squared error, which is the loss function for our model. We report both the mean squared error and the root mean squared error. The results of our fine-tuning are shown in Table 5. As expected, both models perform very well in predicting the grades for reviews in the test split, with the large model performing slightly better.

5. Bias evaluation

We run several tests that evaluate both intrinsic and extrinsic biases to explore possible biases that may have been learned by training models on PRW data. The tests need to be appropriate for German because our data only contains German. Therefore, we use existing metrics and adapt them to the German language. We then use them to evaluate our models in Section 5.2. Then, in Section 5.3, we explore the bias in the two downstream tasks, ABSA and grade prediction, presented in Section 4.1 and Section 4.2, respectively.

5.1. Correlation in data

This study was motivated by the confirmed hypothesis that gender bias exists in the SPAC data. This bias was found at the level of the physician and did not take into account the individual text reviews and their corresponding ratings [2]. Since we use the SPAC data to predict ratings on the basis of individual text reviews and to extract aspect phrases and their polarity, we are interested in a relationship between sensitive attributes such as gender and migration background and rating.

We use correlation analysis to measure the relationship. For this analysis, we use only the filtered data, that we described in Section 3.2. We compute the Spearman's ρ for the relationships between the overall grade of a review and the physician's gender and the migration background, respectively. The Spearman's ρ effect size is 0.013 for the relationship between the overall grade and the gender of the physician. For the relationship between overall grade and migration background, the effect size is −0.022. Both results indicate a very low correlation between the sensitive attribute and the given grade. Thus, no clear relationship can be observed.

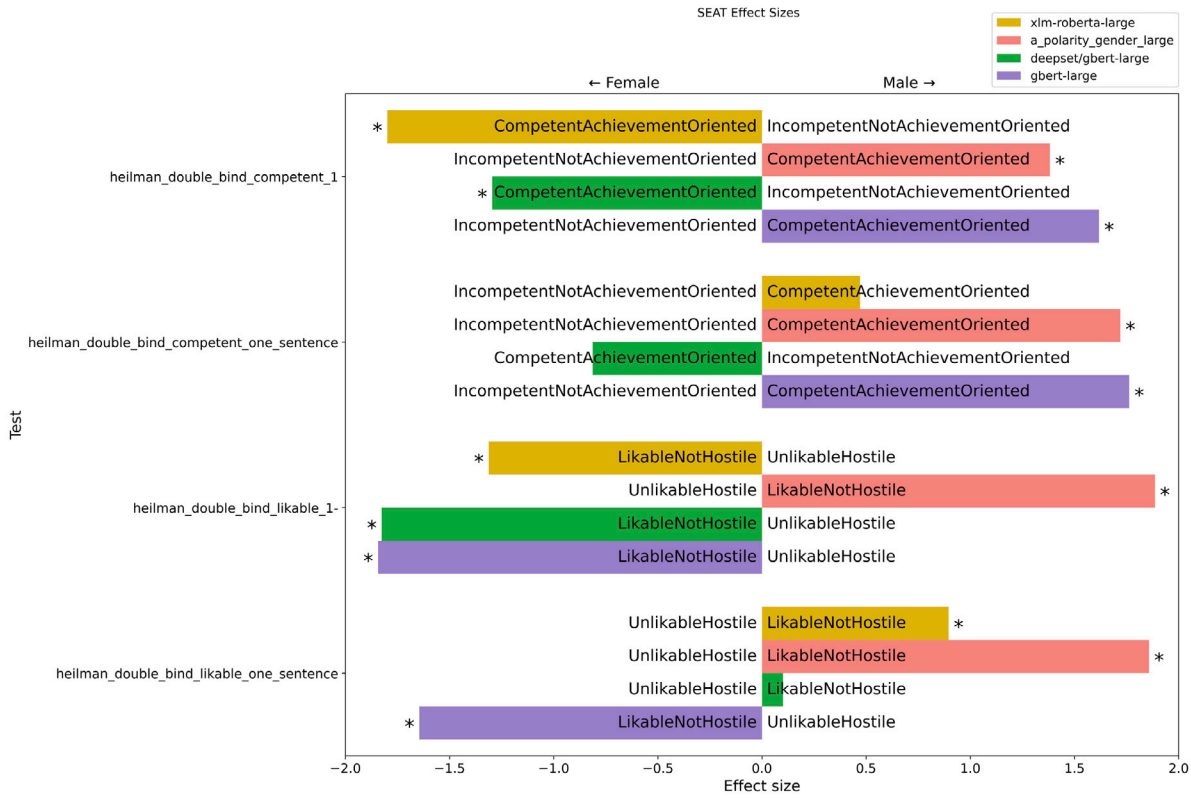


Fig. 5. SEAT results for double-bind tests translated to German. Asterisks indicate a p value below 0.05 and thus statistical significance.

5.2. Intrinsic bias

As shown in Section 2.3.1, intrinsic bias can be explored using different techniques. For our study, we choose two widely used approaches, namely SEAT in Section 5.2.1 and Crows-Pairs in Section 5.2.2, translate the gender-specific tests into German, and evaluate our models.

5.2.1. SEAT

SEAT extends the work of WEAT [27] by using the word sets and filling them into sentence templates. For example, in WEAT. the names *John* and *Amy* are defined for the target categories *MaleNames* and *FemaleNames*, respectively. SEAT reuses these categories and names, and fills them into simple templates, resulting in *This is John* and *This is Amy*. The sentence templates are only available in English [20]. Since we train the models on German-language tasks, we translate selected tests into German. Our selection depends on the gender bias we want to evaluate. Therefore, we use the double-bind and gender-specific SEAT templates (sent-weat 6, 6b, 7, 8, and 8b), translate them using machine translation, and check their correctness manually. We evaluate all of our fine-tuned models for ABSA and grade prediction. Since the models are based on open-source pre-trained parameters, we also run the tests on the pre-trained models to compare them with the fine-tuned models.

WEAT defines a test statistic that measures the differential association of two sets of target words with two sets of attribute words. The effect size is a normalized measure of this test statistic. The p-value is one-sided and indicates the probability of the null hypothesis that a randomly permuted sample would produce a test statistic as large as the observed one [27]. These measures are reused in SEAT [20]. We are interested in the probability that a randomly permuted sample would produce a test statistic as extreme as the observed one under the null hypothesis. That is, for negative (positive) values of the test statistic, we want to know the probability that a randomly permuted sample would have an effect size as small (large) as the observed one under the null hypothesis. Therefore, for negative effect sizes, we report p values as $1 - p$.

Fig. 5 shows the results for the German-translated double-bind tests. We plot the results of the pre-trained and fine-tuned models for grade prediction and ABSA. The pre-trained *xlm-roberta-large* model was fine-tuned for ABSA and the *deepset/gbert-large* model was fine-tuned for grade prediction. The bars show the effect size for a model in each double bind test. For *heilman_double_bind_component_1* the effect size is −1.8, indicating a high relative association of target *Female* with attribute *CompetentAchievementOriented* compared to target *Male*. The asterisk indicates a p-value below .05 and thus a statistical significance. In almost all tests, the sign of the effect size changes or the absolute value increases after fine-tuning. The first two tests show an

Table 6
Results for the gender-subset of CrowS-Pairs translated to German.

Model	Metric score	Stereotype score	Anti-stereotype score
deepset/gbert-large	54.96%	60.13%	47.57%
grade-prediction/gbert-large	52.67%	64.78%	33.98%
xlm-roberta-large	48.85%	48.10%	50.49%
a_polarity_gender_large	47.71%	46.54%	49.51%
b_polarity_gender_large	53.05%	67.30%	31.07%
c_polarity_gender_large	50.38%	45.28%	58.25%
d_polarity_gender_large	57.63%	53.46%	64.08%

association of the fine-tuned models of the target *Male* with the attribute *CompetentAchievementOriented*. This in turn implies a higher relative association of *Female* with *IncompetentAchievementOriented*. These results may indicate a gender bias that degrades women compared to men.

The gender tests, which attribute the likability to genders do not show such a clear pattern. While the fine-tuned grade prediction model has learned an association of *Male* with *LikableNotHostile*, the ABSA model learns a strong association of *Female* with *LikableNotHostile*. Overall, almost all tests are statistically significant and show a high absolute effect size. This implies a strong difference in the association between the two sets of target and attribute sentences. Further results of the German-translated SEAT templates of the best performing models can be found in the Appendix in Table 9. As with the tests discussed before, the results do not show a clear pattern of unidirectional bias toward a single gender.

5.2.2. CrowS-Pairs

As a second intrinsic bias measure, we perform an evaluation using the CrowS-Pairs [23]. The underlying data is only available in English, but the technique seems to be applicable in German as well. Therefore, we translate the sentences corresponding to the gender tests into German. We first evaluate the pre-trained models *xlm-roberta-large* and *deepset/gbert-large* before running the tests on our fine-tuned versions for ABSA and grade prediction.

The scale for the CrowS metric ranges from 0 to 100. It is a weighted average of the stereotype score and the anti-stereotype score. The stereotype score measures the percentage of examples, where a model would prefer to generate the more stereotypical sentence. The anti-stereotype score measures the percentage of examples, where a model would prefer to generate the more anti-stereotypical sentence. Thus, a model treats a bias category in the most fair way, if the score is at 50%.

The results for the gender-specific German CrowS-Pairs are shown in Table 6. For *deepset/gbert-large*, a moderate score of 54.96% is reduced by fine-tuning the grade prediction, although this is achieved by a higher stereotype score combined with a lower anti-stereotype score. In addition, *xlm-roberta-large*, fine-tuned to ABSA, shows all kinds of changes in the metric, as the score is reduced for dataset a, is nearly perfect for c, and increases for b and c. Although almost all of the resulting scores indicate an unfairness in the decision between stereotyping and anti-stereotyping judgments toward gender, no clear direction of bias can be observed.

One aspect that the CrowS-Pairs does not measure is language modeling ability. Pre-trained models are usually trained for MLM. When fine-tuning the model for a downstream task, e.g., ABSA or grade prediction, the MLM layer is removed and another layer is added, e.g., for regression or token classification. The new layer is then fine-tuned along with the parameters for the embedding layer. When the fine-tuned model is reused for MLM, the last layer must be replaced by a new MLM layer with randomly initialized parameters that do not match the parameters of the embedding layer. Therefore, without retraining for this task, the language modeling capability may suffer. For this reason, the results of MLM-based fairness metrics should be interpreted with caution. Since the MLM layer is randomly initialized, the results may also be partially random. Therefore, the probabilities for certain target or attribute words may not represent the associations that a model with adequate language modeling ability would generate.

5.3. Extrinsic bias

The measurement of extrinsic bias depends on the downstream tasks being tested. Therefore, in Section 5.3.1 we evaluate the performance of the models tuned to ABSA, a token classification task while attributing for sensitive attributes. In Section 5.3.2, we evaluate the grade prediction model that has been fine-tuned for a regression task, also with sensitive group attribution.

5.3.1. ABSA

Our approach to ABSA involves using a single model to extract aspect terms, aspect categories, and polarities of extracted terms. This allows direct analysis of the learned bias. By outputting the polarity of extracted and categorized terms, the performance of a model in classifying positive or negative aspect terms can be related to sensitive attributes. That is, for both women and men or migrants and natives, a fair model would have the same probability and error rate of classifying a term as positive or negative. Therefore, we compute parity-based fairness measures to investigate whether our fine-tuned ABSA models are unfair to the gender or migration background of the rated physicians. For each possible label, we compute the group-wise accuracy, precision, recall, and F1 score for the gender (female and male) and the migration background (migrant and native). The results of this evaluation for the best performing xlm-roberta-large model on dataset A are shown in Table 7. The results for the best performing models on the other datasets can be found in the Appendix.

Table 7
Group-based metrics for XLM-RoBERTa-large on the gender-annotated subset of dataset A.

Metric	Category	Polarity	Female	Male	Migrant	Native	Overall
Accuracy	Macro		0.87	**0.88**	**0.88**	**0.88**	0.88
F1	Macro		**0.59**	0.57	0.50	**0.60**	0.60
	Explanation	P	0.63	**0.72**	0.59	**0.72**	**0.70**
		N	**0.54**	0.31	0.00	**0.49**	0.49
	Friendliness	P	0.72	**0.73**	**0.78**	0.72	**0.73**
		N	0.52	**0.67**	0.62	**0.63**	0.63
	Competence	P	**0.79**	0.69	**0.73**	0.72	**0.72**
		N	**0.39**	0.24	0.00	**0.36**	0.34
	Time Taken	P	0.78	**0.80**	**0.81**	0.79	**0.79**
		N	0.00	**0.05**	0.00	**0.04**	0.04
	O-tag		**0.93**	**0.93**	**0.93**	**0.93**	0.93
Precision	Macro		0.66	**0.73**	0.55	**0.73**	0.72
	Explanation	P	0.52	**0.67**	**0.71**	0.63	0.64
		N	**0.87**	0.67	0.00	**0.83**	**0.83**
	Friendliness	P	0.68	**0.77**	**0.78**	0.72	**0.73**
		N	0.47	**0.70**	**0.80**	0.61	0.62
	Competence	P	**0.80**	0.70	**0.79**	0.72	0.73
		N	0.80	**1.00**	0.00	**0.84**	**0.84**
	Time Taken	P	**0.88**	0.82	**0.91**	0.83	**0.84**
		N	0.00	**0.33**	0.00	**0.50**	0.33
	O-tag		**0.93**	0.92	0.91	**0.93**	0.93
Recall	Macro		**0.58**	0.54	0.46	**0.58**	0.57
	Explanation	P	**0.78**	0.77	0.51	**0.86**	**0.77**
		N	**0.39**	0.20	0.00	**0.35**	0.35
	Friendliness	P	**0.77**	0.69	**0.77**	0.71	**0.72**
		N	0.58	**0.65**	0.50	**0.64**	0.63
	Competence	P	**0.77**	0.68	0.68	**0.72**	**0.71**
		N	**0.26**	0.13	0.00	**0.23**	0.21
	Time Taken	P	0.69	**0.78**	0.73	**0.76**	**0.75**
		N	0.00	**0.03**	0.00	**0.02**	0.02
	O-tag		**0.94**	**0.94**	**0.96**	0.94	0.94

The accuracy over all predicted tokens is about 0.88. The differences between the groups are negligible. Although precision is better for the negative categories *explanation* and *competence*, group-independent recall and F1 values are higher for all positive categories compared to negative categories. The imbalanced overall performance in terms of term polarity may result from the highly imbalanced distribution of positive and negative terms in the training and test subset. The distribution within the test subset is shown in Table 2 and shows that there are significantly fewer negative aspect terms than positive ones. This is also consistent with the relative distribution of physician grades, which has a large proportion of very positive grades, as shown in Fig. 1.

There are significant differences in scores in both directions for all four groups when examining the group-wise performance of the model for categories within a polarity. If the data is not balanced, it could also influence the result. In other words, the labels are not evenly spread across sexes or ethnic groups. Therefore, there may be large discrepancies between the performance of the groups. For example, the *explanation-N* and *competence-N* classes within the *migrant* and *native* groups contain either no tokens or very few tokens in the test dataset, which can lead to values of 0.

For positive categories, the results appear to be relatively consistent. The differences in F1 values within groups range from 0.01 to a maximum of 0.1. This suggests that the model is impartial in assessing physicians of different genders and immigrant backgrounds, as no group shows a clear disadvantage in positive aspects. This conclusion cannot be drawn for negative aspects, as there are large differences in performance between all groups. Nor is there any clear performance disadvantage for any single group. Thus, these differences may be due to the underrepresentation of negative items within groups rather than to an inherent unfairness in the model.

5.3.2. Grade prediction

For the grade prediction task, we first evaluate the correlation of the predicted grade with the sensitive attributes, analogous to Section 5.1. We use our fine-tuned *xlm-roberta-large* model to predict the overall grades of individual text reviews from the test data. The Spearman's ρ effect size for the correlation between gender and predicted grade is 0.002. For the correlation between migration background and predicted grade, the effect size is −0.023. Consistent with the correlation analysis within the data, these results do not indicate a relationship between the sensitive attributes and the grades predicted by the model.

Table 8
Group-based metrics for deepset/gbert-large fine-tuned on grade prediction.

Metric	Female	Male	Migrant	Native	Overall
MSE	0.067	0.066	0.067	0.066	0.066
RMSE	0.259	0.257	0.258	0.257	0.257

Table 9
SEAT results on gender-specific templates for the best-performing models for our tasks.

Test	xlm-roberta-large	a_polarity_gender_large	b_polarity_gender_large	c_polarity_gender_large	d_polarity_gender_large	deepset/gbert-large	grade-prediction
Double Bind: Competent	−1.80*	1.38*	1.91*	−1.91*	1.74*	−1.29*	1.62*
Double Bind: Competent - sent	−1.80	1.72*	1.91	−1.91	1.54*	−1.29	1.76*
Double Bind: Likable	−1.31*	1.89*	−1.56*	−1.72*	1.80*	−1.83*	−1.84*
Double Bind: Likable - sent	0.90*	1.86*	−1.56	1.39*	1.24*	−1.83	−1.65*
sent-weat6	0.90	1.86	0.58*	0.49*	0.39*	0.10	−0.38*
sent-weat6b	−0.23	−0.08	0.75*	0.49	0.36*	0.17	−0.68*
sent-weat7	−0.46*	−0.78*	0.65*	−0.85*	0.36	−0.19	−0.57*
sent-weat8	−0.46	0.68*	−0.46*	−0.85	−0.45*	0.06	−0.57
sent-weat8b	−0.27	0.68	−0.46*	−0.28	−0.45*	0.02	−0.04

Asterisks indicate ρ-values lower than 0.05.

The group-wise evaluation of performance metrics that we did for ABSA in Section 5.3.1 is also applicable to the grade prediction task. Since grade prediction is a regression task, we use the mean squared error and the root mean squared error computed on the test split for each sensitive group. The results of our best xlm-roberta-large model are shown in Table 8 and show no meaningful differences in the errors for any of the groups. The underlying distribution of the groups in the test split is unbalanced, as there are 186,188 ratings of male physicians and 94,349 ratings of female physicians. The background is also unbalanced with 246,136 native physicians and 34,401 immigrant physicians.

6. Correlation analyses

To evaluate whether the results of the intrinsic bias measures have an effect on the extrinsic bias measures, we perform correlation analyses. For each model that we fine-tuned in Section 4, we have computed SEAT scores in Section 5.2.1 and CrowS-Pairs scores in Section 5.2.2. In addition, we computed extrinsic, group-based performance measures for all fine-tuned models in Section 5.3. The availability of this set of metric results allows for an analysis of the influence of intrinsic metric results on extrinsic metric results.

The first analysis we perform concerns the effect of SEAT scores on the performance of ABSA models. In Section 4.1, we aggregated the ABSA corpus with the gender and migration attributes of SPAC. To better distinguish the number of possible aspect categories, we divided the corpus into four subsets: A, B, C, and D. Each subset was used to fine-tune XLM-RoBERTa in the base and large versions. The group-wise performance evaluation includes precision, recall, and f1 scores for all categories, polarities, models, and datasets. SEAT is a normalized measure of the difference in the associations of two target sets with two attribute sets [20]. For our study, we have translated gender-specific templates into German. Therefore, the effect size indicates the difference in the relative associations of gender with positive or negative attributes. In the same way that SEAT measures a gender difference in representations, generated by a model, we measure the gender difference in terms of model performance. That is, we compute the difference in macro-averaged F1 scores for each model. This allows us to correlate the SEAT results of each template with the gender difference in F1 scores for each fine-tuned model.

Moreover, in Section 5.2.2 we have evaluated all the fine-tuned models on the CrowS-Pairs benchmark. To do this, we used only gender-specific sentence pairs and translated them into German before calculating the metric scores. The metric scores indicate whether a model tends to predict stereotypical or anti-stereotypical sentences with respect to gender. Thus, a score of 50 would be interpreted as a fair model, and lower or higher scores would indicate a difference in the associations a model shows toward gender. Therefore, we can also correlate the CrowS-Pairs scores with the gender difference in the F1 score for each fine-tuned model. We compute Spearman's ρ for all gender-specific SEAT templates and the gender-specific CrowS-Pairs benchmark and show the results in Fig. 6.

Both the SEAT effect size and the group-wise difference in model performance would be expected to increase if gender bias were present in a fine-tuned model. The same pattern would also be expected for the combination of the CrowS-Pairs benchmark and the group-wise performance difference. Furthermore, this would result in large correlation coefficients. This is because larger (smaller) absolute effect sizes (metric scores) would result in larger (smaller) differences in F1 score. The correlation coefficients that we have calculated do not agree with this expected result for all the combinations. The correlation coefficients for the SEAT

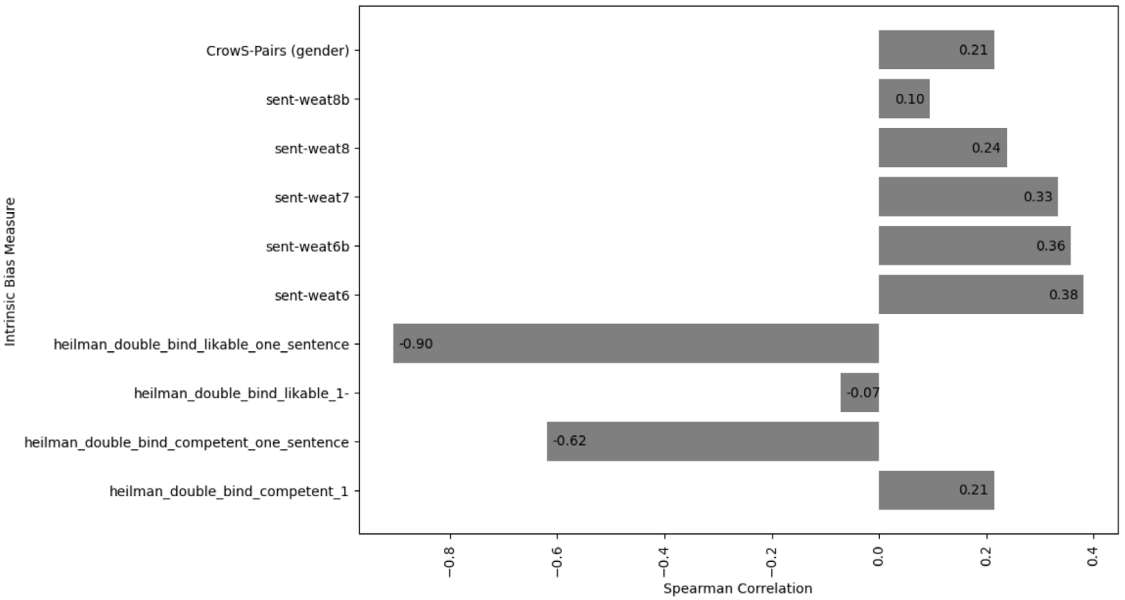


Fig. 6. Spearman correlation of the difference in gender F1 score with the SEAT effect size.

Table 10
Number of tokens per tag in gender-annotated test subset of dataset B.

Tag	Polarity	Male	Female	Native	Migrant	Overall
O		2,794	1,960	4,343	411	4,754
Alternative Healing Methods	P	17	25	38	4	42
	N	14	3	17	0	17
Treatment	P	82	65	126	21	147
	N	20	31	44	7	51
Care/Commitment	P	103	84	163	24	187
	N	2	11	13	0	13
Overall/Recommendation	P	174	179	315	38	353
	N	44	81	113	12	125
Child-friendliness	P	128	48	169	7	176
	N	26	10	36	0	36
Relationship of Trust	P	295	154	405	44	449
	N	10	9	16	3	19

templates sent-weat6, 6b, 7, 8, and 8b, and the gender-specific CrowS-Pairs are all positive values and range from 0.1 to 0.38. This indicates that as the effect size (metric score) increases, the difference in F1 values between female and male physicians increases. In contrast to this observation are the results of the correlation with the double-bind templates. Here, although all double-bind templates assess gender associations, both the polarity and the absolute values of the correlation coefficients change dramatically for different templates.

One thing that needs to be considered is the number of samples used to compute the correlations. For this initial analysis, we had a total of 16 fine-tuned models. Therefore, we had 16 pairs of metric scores and macro F1 differences for each test. Therefore, a larger sample size of additional fine-tuned models would provide more meaningful or more reliable results, since the outliers would have less of a significant impact. Furthermore, due to the uneven distribution of tokens in the training and test splits, it is questionable whether the group-wise performance differences are solely due to a potentially biased model. It is possible that the uneven distribution or the small number of tokens, some of which were significantly small, may have biased the group-wise results in such a way that the inherent bias of a model did not come to light. In addition, future analyses could evaluate the correlations not only on the differences of the macro F1 scores but also on the individual level of the categories for the F1 scores and the precision or recall.

If the results of our first correlation analysis were interpreted in such a way that a large metric score would correlate with a large difference in group-wise model performance on a downstream task, then this correlation would also apply to the grade prediction task, which was presented in Section 4.2. A Spearman's ρ could be computed for the variables' metric score and the group-wise difference in mean squared error. Since we only trained two models on this task due to the immediate strong performance, the

Table 11
Groupwise metrics for XLM-RoBERTa-large on the gender-annotated subset of dataset B.

Metric	Category	Polarity	Female	Male	Migrant	Native	Overall
Accuracy	Macro		0.84	**0.87**	0.85	**0.86**	0.86
F1	Macro		0.42	**0.47**	0.41	**0.45**	0.45
	Alternative Healing Methods	P	**0.85**	0.71	0.00	**0.80**	**0.78**
		N	0.00	0.00	0.00	0.00	0.00
	Treatment	P	**0.69**	0.65	**0.67**	**0.67**	**0.67**
		N	0.00	0.00	0.00	0.00	0.00
	Care/Commitment	P	0.51	**0.56**	**0.76**	0.50	**0.54**
		N	0.00	0.00	0.00	0.00	0.00
	Overall/Recommendation	P	0.63	**0.72**	**0.73**	0.67	**0.68**
		N	**0.41**	0.38	**0.69**	0.36	0.40
	Child-Friendliness	P	0.63	**0.71**	**1.00**	0.67	**0.68**
		N	0.10	**0.41**	0.00	**0.32**	0.32
	Relationship Of Trust	P	0.71	**0.78**	0.63	**0.77**	**0.75**
		N	0.00	0.27	0.00	**0.14**	0.13
	O		0.91	**0.93**	0.91	**0.92**	0.92
Precision	Macro		0.44	**0.57**	0.42	**0.50**	0.50
	Alternative Healing Methods	P	**0.92**	0.55	0.00	**0.70**	**0.70**
		N	0.00	0.00	0.00	0.00	0.00
	Treatment	P	**0.61**	0.55	0.50	**0.59**	**0.57**
		N	0.00	0.00	0.00	0.00	0.00
	Care/Commitment	P	0.63	**0.74**	**0.92**	0.65	**0.69**
		N	0.00	0.00	0.00	0.00	0.00
	Overall/Recommendation	P	**0.68**	0.66	**0.83**	0.65	0.67
		N	0.62	0.79	0.61	0.69	**0.68**
	Child-Friendliness	P	0.57	**0.74**	**1.00**	0.66	**0.68**
		N	0.13	**0.69**	0.00	**0.50**	0.50
	Relationship Of Trust	P	0.75	**0.84**	0.70	**0.82**	**0.81**
		N	0.00	**1.00**	0.00	**0.33**	0.33
	O		0.88	**0.91**	**0.90**	**0.90**	0.90
Recall	Macro		0.41	**0.47**	0.43	**0.44**	0.44
	Alternative Healing Methods	P	0.79	**1.00**	0.00	**0.94**	**0.87**
		N	0.00	0.00	0.00	0.00	0.00
	Treatment	P	**0.81**	**0.81**	**1.00**	0.78	**0.81**
		N	0.00	0.00	0.00	0.00	0.00
	Care/Commitment	P	0.42	**0.46**	**0.65**	0.41	**0.44**
		N	0.00	0.00	0.00	0.00	0.00
	Overall/Recommendation	P	0.58	**0.79**	0.65	**0.69**	**0.69**
		N	**0.30**	0.25	**0.79**	0.24	0.28
	Child-Friendliness	P	**0.70**	0.69	**1.00**	0.67	**0.69**
		N	0.08	**0.29**	0.00	**0.24**	0.24
	Relationship Of Trust	P	0.68	**0.72**	0.58	**0.73**	**0.71**
		N	0.00	**0.15**	0.00	**0.09**	0.08
	O		0.94	**0.95**	0.93	**0.95**	0.95

sample size for each intrinsic bias measure is 2. This does not allow for a meaningful correlation analysis, since all coefficients would be either 1 or −1. The group-wise differences in mean squared error for the two-grade prediction models are at 0.00098 and 0.00095 for the base and large models, respectively. If these small differences were repeated for additional fine-tuned models, low correlation coefficients would be expected.

7. Conclusion

In this study, we evaluated how biased data affects both intrinsic and extrinsic language model bias metrics for various downstream tasks by combining different datasets, training approaches, and bias measurement techniques.

Based on three corpora (PRW collection, SPAC, and ABSAC), we have defined two downstream tasks, namely ABSA and grade prediction. For ABSA, we proposed a new approach that combines the three tasks ATE, ACC, and APC. Due to the reduced data availability for APC and the gender and migration attributes, that were aggregated with the ABSAC, the performance of our

Table 12
Number of tokens per tag in gender-annotated test subset of dataset C.

Tag	Polarity	Male	Female	Native	Migrant	Overall
O		3,631	1,844	4,870	605	5,475
Care/Commitment	P	105	85	166	24	190
	N	0	18	7	11	18
Friendliness	P	232	102	305	29	334
	N	64	47	105	6	111
Competence	P	129	45	159	15	174
	N	26	52	76	2	78
Accessibility by Telephone	P	47	50	82	15	97
	N	197	156	330	23	353

Table 13
Groupwise metrics for XLM-RoBERTa-large on the gender-annotated subset of dataset C.

Metric	Category	Polarity	Female	Male	Migrant	Native	Overall
Accuracy	Macro		0.85	**0.89**	0.87	**0.88**	0.88
F1	Macro		**0.57**	**0.57**	0.54	**0.57**	0.57
	Care/Commitment	P	**0.42**	0.38	0.28	**0.42**	**0.40**
		N	0.00	0.00	0.00	0.00	0.00
	Friendliness	P	0.74	**0.79**	0.70	**0.78**	**0.78**
		N	**0.69**	0.59	0.00	**0.64**	0.63
	Competence	P	**0.70**	0.57	**0.74**	0.59	**0.61**
		N	0.33	**0.42**	**0.89**	0.34	0.37
	Accessibility By Telephone	P	0.59	**0.66**	0.39	**0.66**	0.63
		N	0.70	**0.78**	**0.92**	0.74	**0.75**
	O		0.92	**0.95**	0.93	**0.94**	0.94
Precision	Macro		**0.61**	0.60	0.54	**0.61**	0.60
	Care/Commitment	P	**0.40**	0.38	0.21	**0.43**	**0.39**
		N	0.00	0.00	0.00	0.00	0.00
	Friendliness	P	0.70	**0.74**	0.63	**0.74**	**0.73**
		N	**0.75**	0.66	0.00	**0.70**	0.69
	Competence	P	**0.80**	0.68	**0.73**	0.71	**0.71**
		N	0.58	**0.62**	**1.00**	0.57	0.60
	Accessibility By Telephone	P	0.59	**0.64**	0.44	**0.64**	0.62
		N	0.75	**0.76**	**0.96**	0.75	**0.76**
	O		0.90	**0.95**	**0.94**	0.93	0.93
Recall	Macro		0.55	**0.56**	**0.55**	**0.55**	0.55
	Care/Commitment	P	**0.45**	0.38	0.40	**0.41**	**0.41**
		N	0.00	0.00	0.00	0.00	0.00
	Friendliness	P	0.78	**0.85**	0.80	**0.84**	**0.83**
		N	**0.65**	0.53	0.00	**0.60**	0.58
	Competence	P	**0.62**	0.49	**0.76**	0.51	**0.53**
		N	0.23	**0.32**	**0.80**	0.25	0.27
	Accessibility By Telephone	P	0.59	**0.69**	0.35	**0.69**	0.64
		N	0.65	**0.81**	**0.89**	0.73	**0.74**
	O		0.94	**0.95**	0.93	**0.95**	0.95

fine-tuned models was reasonable but did not exceed previous modeling approaches. For the grade prediction task, our results immediately showed a strong performance in terms of mean squared error. Aggregating SPAC with the ABSAC for ABSA and with the PRW collection added gender and migration attributes to all training and test examples in the datasets used for the defined tasks. This allowed us to analyze the effect of gender on bias measures and, where possible, the effect of migration bias.

For this reason, we evaluated the bias metrics on our classification and regression tasks. The first measures we computed are gender-specific SEAT [20] templates that we translated into German. The results in Section 5.2.1 show different outputs for the same target-attribute combinations in different SEAT templates. Although the context sentences vary in these templates, the targets and attributes cover the same categories, so unidirectional SEAT scores were expected to result for all gender-specific tests of a model. Complementing our evaluation of SEAT templates, we performed evaluations of gender-specific CrowS-Pairs [23] in Section 5.2.2, which we also translated into German. The computed metric scores did not show a clear pattern of unidirectional unfairness. All evaluated models produced different results, favoring both stereotypes and anti-stereotypes in gender-specific sentence pairs.

Table 14
Number of tokens per tag in gender-annotated test subset of dataset D.

Tag	Polarity	Male	Female	Native	Migrant	Overall
O		2,607	1,706	3,784	529	4,313
Equipment	P	46	28	50	24	74
	N	1	0	1	0	1
Waiting time in the practice	P	123	101	212	12	224
	N	85	64	127	22	149
Waiting Time for an appointment	P	197	172	318	51	369
	N	50	22	62	10	72
Well Being	P	350	201	482	69	551
	N	75	13	74	14	88

Table 15
Groupwise metrics for XLM-RoBERTa-large on the gender-annotated subset of dataset D.

Metric	Category	Polarity	Female	Male	Migrant	Native	Overall
Accuracy	Macro		0.88	**0.89**	**0.91**	0.89	0.89
F1	Macro		0.67	**0.69**	**0.76**	0.67	0.68
	Equipment	P	**0.82**	0.66	**0.82**	0.69	**0.73**
		N	0.00	0.00	0.00	0.00	0.00
	Waiting Time In The Practice	P	0.64	**0.75**	0.68	**0.71**	**0.71**
		N	**0.67**	0.65	**0.87**	0.62	0.66
	Waiting Time For An Appointment	P	0.83	**0.84**	**0.83**	**0.83**	0.83
		N	0.76	**0.89**	**0.96**	0.83	**0.84**
	Well Being	P	0.77	**0.79**	0.78	**0.78**	**0.78**
		N	0.63	**0.67**	**1.00**	0.58	0.66
	O		0.93	**0.94**	**0.94**	**0.94**	0.94
Precision	Macro		0.64	**0.68**	**0.76**	0.65	0.67
	Equipment	P	0.69	**0.70**	**0.72**	0.68	**0.69**
		N	0.00	0.00	0.00	0.00	0.00
	Waiting Time In The Practice	P	0.71	**0.79**	0.70	**0.77**	**0.76**
		N	0.56	**0.59**	**0.77**	0.54	0.58
	Waiting Time For An Appointment	P	**0.91**	0.84	**0.96**	0.86	**0.87**
		N	0.66	**0.80**	**0.92**	0.73	0.75
	Well Being	P	**0.84**	0.82	0.79	**0.83**	**0.82**
		N	0.49	**0.68**	**1.00**	0.54	0.62
	O		0.93	**0.94**	**0.94**	0.93	0.93
Recall	Macro		**0.73**	0.70	**0.78**	0.69	0.71
	Equipment	P	**1.00**	0.63	**0.94**	0.70	**0.77**
		N	0.00	0.00	0.00	0.00	0.00
	Waiting Time In The Practice	P	0.59	**0.72**	**0.67**	0.66	0.66
		N	**0.83**	0.73	**1.00**	0.73	**0.77**
	Waiting Time For An Appointment	P	0.76	**0.83**	0.73	**0.81**	0.80
		N	0.90	**1.00**	**1.00**	0.96	**0.97**
	Well Being	P	0.72	**0.76**	**0.77**	0.75	**0.75**
		N	**0.88**	0.66	**1.00**	0.63	0.70
	O		0.93	**0.94**	**0.94**	**0.94**	0.94

Moreover, the reliability of this benchmark on our models is questionable since they were fine-tuned for different downstream tasks without any training for the necessary MLM task required by this benchmark. Therefore, the reliability should be further investigated by combining the benchmark with measuring the language modeling ability, exemplified by the approach taken in StereoSet [31]. The extrinsic measures that we applied to our fine-tuned models are group-wise performance measures that measure the performance of a model on the subset of data within a sensitive group and compare the results. Therefore, for our approach to ABSA, we have evaluated the accuracy, precision, recall, and F1 score for all predictable categories in the sensitive groups of male, female, migration background, and native background. The results of the extrinsic bias measurements on the ABSA task in Section 5.3.1 did not show clear patterns of unidirectional gender bias. Performance varied at both the macro-average and individual category levels, favoring female or male or native or immigrant physicians for different categories. We conclude that these performance differences for different sensitive groups may be due to the unbalanced data distribution since performance for sufficiently available categories does not show significant performance differences. In Section 5.3.2, performance on the grade prediction task within the sensitive

groups also showed no significant differences. This supports the hypothesis that enough training data, even if unbalanced, leads to fair model performance with respect to sensitive groups.

Therefore, it is not possible to conclude that gender-biased data will inherently lead to biased models. The impact of such biased data depends on many variables, such as data volume, data distribution, modeling approach, scenario, and metrics, and must be assessed using multiple appropriate metrics. It is not enough to rely on selected metrics. Intrinsic measures do not necessarily correlate with the fairness of a model in a downstream task, but they can have an impact on extrinsic measures, as pointed out in Section 6. In addition, most intrinsic measures depend on templates, which are always subsets of all possible entries for a given category. For example, templates containing women's names can never cover all possible women's names that exist. As a result, measuring bias using the set of templates will focus only on the names that are defined. The use of other names that are not part of the tested template set could have a significant impact on the results. In addition, even two sentences with the same semantic meaning but different structures or wording could result in different metric scores. Conclusions drawn from one set of templates cannot be generalized to other templates or to effects on extrinsic measures. Finally, the use of fairness measures leaves a lot of room for interpretation of the results. Therefore, depending on the context in which a model is to be trained and used, multiple measures must be interpreted, prioritized, and weighted. That is, more or less strict limits or rules must be applied depending on the real-world implications that unfairness has with respect to a chosen metric.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was co-funded by the German Federal Ministry of Education and Research under the grant 13N16242.

Appendix

See Tables 9–15.

References

[1] J. Kersting, M. Geierhos, Towards aspect extraction and classification for opinion mining with deep sequence networks, in: R. Loukanova (Ed.), Natural Language Processing in Artificial Intelligence, NLPinAI 2020, in: Studies in Computational Intelligence (SCI), vol. 939, Springer, Cham, Switzerland, 2021, pp. 163–189, http://dx.doi.org/10.1007/978-3-030-63787-3_6.

[2] M. Kauff, J. Anslinger, O. Christ, M. Niemann, M. Geierhos, L. Huster, Ethnic and gender-based prejudice towards medical doctors? The relationship between physicians' ethnicity, gender, and ratings on a physician rating website, J. Soc. Psychol. 162 (5) (2022) 540–548, http://dx.doi.org/10.1080/00224545.2021.1927944.

[3] M. Emmert, F. Meier, An analysis of online evaluations on a physician rating website: Evidence from a german public reporting instrument, J. Med. Internet Res. 15 (8) (2013) e157, http://dx.doi.org/10.2196/jmir.2655.

[4] M. Emmert, F. Meier, F. Pisch, U. Sander, Physician choice making and characteristics associated with using physician-rating websites: Cross-sectional study, J. Med. Internet Res. 15 (8) (2013) e187.

[5] C. Ellimoottil, S.W. Leichtle, C.J. Wright, A. Fakhro, A.K. Arrington, T.J. Chirichella, W.H. Ward, Online physician reviews: The good, the bad and the ugly, Bull. Am. College Surgeons 98 (9) (2013) 34–39.

[6] M. Emmert, U. Sander, F. Pisch, Eight questions about physician-rating websites: A systematic review, J. Med. Internet Res. 15 (2) (2013) e24, http://dx.doi.org/10.2196/jmir.2360.

[7] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013, URL http://arxiv.org/abs/1301.3781.

[8] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: EMNLP, Association for Computational Linguistics, 2014, pp. 1532–1543, http://dx.doi.org/10.3115/v1/D14-1162.

[9] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. ACL 5 (2017) 135–146.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st Conference on Neural Information Processing Systems, Curran Associates, Long Beach, CA, USA, 2017, pp. 5998–6008.

[11] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal sentence encoder for english, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2018, pp. 169–174, http://dx.doi.org/10.18653/v1/D18-2029.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, URL https://ui.adsabs.harvard.edu/abs/2019arXiv190711692L.

[14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the ACL, ACL, Online, 2020, pp. 8440–8451, http://dx.doi.org/10.18653/v1/2020.acl-main.747.

[15] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020, URL https://openreview.net/forum?id=H1eA7AEtvS.

[16] K. Clark, M.-T. Luong, Q.V. Le, C.D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020, URL https://openreview.net/forum?id=r1xMH1BtvB.
[17] P. Delobelle, E. Tokpo, T. Calders, B. Berendt, Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2022, pp. 1693–1706, http://dx.doi.org/10.18653/v1/2022.naacl-main.122.
[18] S. Zhang, P. Li, Z. Cai, Are male candidates better than females? Debiasing BERT resume retrieval system, in: 2022 IEEE International Conference on Systems, Man, and Cybernetics, SMC, 2022, pp. 616–621, http://dx.doi.org/10.1109/SMC53654.2022.9945184.
[19] M. Bartl, M. Nissim, A. Gatt, Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias, in: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, 2020, pp. 1–16, URL https://aclanthology.org/2020.gebnlp-1.1.
[20] C. May, A. Wang, S. Bordia, S.R. Bowman, R. Rudinger, 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), Association for Computational Linguistics, 2019, pp. 622–628, http://dx.doi.org/10.18653/v1/N19-1063.
[21] K. Kurita, N. Vyas, A. Pareek, A.W. Black, Y. Tsvetkov, Measuring bias in contextualized word representations, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, 2019, pp. 166–172, http://dx.doi.org/10.18653/v1/W19-3823.
[22] K. Webster, X. Wang, I. Tenney, A. Beutel, E. Pitler, E. Pavlick, J. Chen, S. Petrov, Measuring and reducing gendered correlations in pre-trained models, 2020, CoRR abs/2010.06032, URL https://arxiv.org/abs/2010.06032.
[23] N. Nangia, C. Vania, R. Bhalerao, S.R. Bowman, Crows-pairs: A challenge dataset for measuring social biases in masked language models, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 1953–1967, http://dx.doi.org/10.18653/v1/2020.emnlp-main.154.
[24] S. Goldfarb-Tarrant, R. Marchant, R. Muñoz Sánchez, M. Pandya, A. Lopez, Intrinsic bias metrics do not correlate with application bias, 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), Association for Computational Linguistics, 2021, pp. 1926–1940, http://dx.doi.org/10.18653/v1/2021.acl-long.150.
[25] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2013, pp. 746–751, URL https://aclanthology.org/N13-1090.
[26] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, A. Kalai, Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, 2016, pp. 4356–4364.
[27] A. Caliskan, J.J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186, http://dx.doi.org/10.1126/science.aal4230.
[28] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 2227–2237, http://dx.doi.org/10.18653/v1/N18-1202.
[29] M.E. Heilman, A.S. Wallen, D. Fuchs, M.M. Tamkins, Penalties for success: Reactions to women who succeed at male gender-typed tasks, J. Appl. Psychol. 89 (3) (2004) 416–427, http://dx.doi.org/10.1037/0021-9010.89.3.416.
[30] W. Guo, A. Caliskan, Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases, 2021, pp. 122–133, http://dx.doi.org/10.1145/3461702.3462536.
[31] M. Nadeem, A. Bethke, S. Reddy, StereoSet: Measuring stereotypical bias in pretrained language models, 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), Association for Computational Linguistics, 2021, pp. 5356–5371, http://dx.doi.org/10.18653/v1/2021.acl-long.416.
[32] S. Caton, C. Haas, Fairness in machine learning: A survey, 2020, http://dx.doi.org/10.48550/arXiv.2010.04053.
[33] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender bias in coreference resolution: Evaluation and debiasing methods, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, 2018, pp. 15–20, http://dx.doi.org/10.18653/v1/N18-2003.
[34] R. Rudinger, J. Naradowsky, B. Leonard, B. Van Durme, Gender bias in coreference resolution, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, 2018, pp. 8–14, http://dx.doi.org/10.18653/v1/N18-2002.
[35] K. Sakaguchi, R.L. Bras, C. Bhagavatula, Y. Choi, WinoGrande: An adversarial winograd schema challenge at scale, Commun. ACM 64 (9) (2021) 99–106, http://dx.doi.org/10.1145/3474381.
[36] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A.T. Kalai, Bias in bios: A case study of semantic representation bias in a high-stakes setting, 2019, pp. 120–128, http://dx.doi.org/10.1145/3287560.3287572.
[37] S. Mishra, S. He, L. Belli, Assessing demographic bias in named entity recognition, 2020, http://dx.doi.org/10.48550/arXiv.2008.03415.
[38] D. Saunders, B. Byrne, 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, Association for Computational Linguistics, 2020, pp. 7724–7736, http://dx.doi.org/10.18653/v1/2020.acl-main.690.
[39] E. Sheng, K.-W. Chang, P. Natarajan, N. Peng, Societal biases in language generation: Progress and challenges, 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), Association for Computational Linguistics, 2021, pp. 4275–4293, http://dx.doi.org/10.18653/v1/2021.acl-long.330.
[40] T. Sun, J. He, X. Qiu, X. Huang, BERTScore is unfair: On social bias in language model-based metrics for text generation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2022, pp. 3726–3739, http://dx.doi.org/10.18653/v1/2022.emnlp-main.245.
[41] J. Kersting, M. Geierhos, Aspect phrase extraction in sentiment analysis with deep learning, in: Proceedings of the 12th International Conference on Agents and Artificial Intelligence: Special Session on Natural Language Processing in Artificial Intelligence, SCITEPRESS, Valetta, Malta, 2020, pp. 391–400.
[42] J. Kersting, M. Geierhos, Neural learning for aspect phrase extraction and classification in sentiment analysis, in: Proceedings of the 33rd International Florida Artificial Intelligence Research Symposium (FLAIRS) Conference, AAAI, North Miami Beach, FL, USA, 2020, pp. 282–285.
[43] J. Kersting, M. Geierhos, Human language comprehension in aspect phrase extraction with importance weighting, in: E. Métais, F. Meziane, H. Horacek, E. Kapetanios (Eds.), Natural Language Processing and Information Systems, in: LNCS, vol. 12801, Springer, Saarbrücken, Germany, 2021, pp. 231–242, http://dx.doi.org/10.1007/978-3-030-80599-9_21.
[44] J. Kersting, M. Geierhos, Well-being in plastic surgery: Deep learning reveals patients' evaluations, in: Proceedings of the 10th International Conference on Data Science, Technology and Applications, SCITEPRESS, Online, 2021, pp. 275–284.
[45] J. Kersting, Identifizierung quantifizierbarer Bewertungsinhalte und -kategorien mittels Text Mining [Identifying Quantifiable Rating Content and Categories Using Text Mining] (Ph.D. thesis), University of the Bundeswehr Munich, 2023.
[46] J. Kersting, M. Geierhos, Towards comparable ratings: Quantifying evaluative phrases in physician reviews, in: A. Cuzzocrea, O. Gusikhin, S. Hammoudi, C. Quix (Eds.), Data Management Technologies and Applications, Springer Nature Switzerland, 2023, pp. 45–65.

[47] O. De Clercq, E. Lefever, G. Jacobs, T. Carpels, V. Hoste, Towards an integrated pipeline for aspect-based sentiment analysis in various domains, in: Proceedings of the 8th ACL Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, ACL, Kopenhagen, Dänemark, 2017, pp. 136–142, http://dx.doi.org/10.18653/v1/w17-5218.
[48] M. Cordes, Wie bewerten die anderen? Eine übergreifende Analyse von Arztbewertungsportalen in Europa [How do the others rate? An overarching analysis of physician rating portals in Europe], 2018, pp. 1–99.

Joschka Kersting. He received his Ph.D. in computer science from the University of the Bundeswehr Munich in 2023. His research interests include aspect-oriented sentiment analysis and AI applications. In the past, he has demonstrated how machines can infer implicit meaning and combine extracted information with additional knowledge.

Falk Maoro. He received his Master's degree in Business Informatics from the University of Applied Sciences in Bielefeld, Germany. Currently, his work at the Research Institute CODE, University of the Bundeswehr Munich, focuses on text classification in various domains. His goal is to make model decisions understandable and trustworthy for users. In particular, he is interested in natural language processing, model bias, and explainable AI.

Michaela Geierhos. She is CODE's Technical Director and Professor for Data Science at the Institute for Data Security in the Department of Computer Science at the Bundeswehr University Munich, Germany. The focus of her research lies at the intersection of computational linguistics and computer science. In this context, she addresses practical problems in natural language processing.