Treating machine learners more like crowd workers:

Increasing validity and robustness of language models with instructions

01.11.2023

Anonymous

Abstract:

Language models like BERT or GPT are becoming increasingly popular tools for creating measurements from large text corpora. While language models tout higher performance than previous models, there is a relevant concern that they behave like "stochastic parrots", reproducing biased patterns from their training data instead of properly measuring the concept they are intended to measure. These hidden biases can be particularly problematic for comparative social science research, where researchers want to compare different social groups (e.g. countries, parties, milieus) and need models to perform equally well on all group members. Investigating this challenge in a comparative analysis across nine groups, four datasets and three types of models, we fine-tune 312 text classifiers and analyse their robustness against group-specific biases and the validity of their outputs. We find that all types of models are susceptible to learning group-specific language patterns and that fine-tuning on biased data (from one group member) reduces performance on representative test sets (from all group members). On average, however, these effects are surprisingly small. In particular when models receive instructions as an additional input, they become more robust against biases from the fine-tuning data and are best at producing valid measurements across different group members. A language model without instructions (BERT-base) sees its average test-set performance decrease by 1.7% F1 macro when trained on biased data compared to random data. Its probability of making an error on group members it has not seen during training increases from 26% to 31.1%. An instruction-based model (BERT-NLI) sees its performance drop by only 0.4% F1 macro when trained on biased data and its probability of making an error on group members it has not seen during training increases only from 18.5% to 19.3%.

1. Introduction

Do our methods actually measure what we think they measure? This is the fundamental question of measurement validity. We can believe to measure ideology in text, but actually measure incumbency (Hirst et al., 2014), we can believe to measure populism, but actually only identify party names (Jankowski & Huber, 2023), or we can believe to measure a flu outbreak but actually measure unrelated search terms like 'basketball' (Lazer et al., 2014). Substantive conclusions drawn from these invalid measurements can be substantially wrong.

The text-as-data literature is one methodological strand in the social sciences that is actively discussing this challenge of validity (Grimmer & Stewart, 2013). While the Natural Language Processing (NLP) literature is constantly developing new methods that report better and better performance, there is still an open debate to what extent these models can also produce valid measurements (Baden et al., 2022). Some linguists provide additional reasons for scepticism, arguing that language models like BERT or GPT behave like "stochastic parrots" that reproduce (biased) patterns from their training data instead of truly understanding their tasks (Bender et al., 2021). This "parrot" characteristic is particularly problematic for comparative social science research. Social scientists often need to validly measure complex concepts across different social groups such as countries, milieus or languages, while working with real-world training data that is never perfectly representative.

Most empirical research related to this challenge of bias originates from the NLP (fairness) literature, which focuses on bias against social group and robustness against spurious language patterns. Social scientists are mostly interested in language models as a measurement tool and biases become most relevant when they impact measurement validity – an issue for which little empirical research exists. Inspired by both strands of literature, this paper asks: To what extent do group-specific language patterns (biases) impact measurement validity in supervised machine learning? To what extent are different supervised machine learning approaches biased by group-specific language patterns in fine-tuning data?

We start by discussing measurement validity (social science literature) and its link to bias and robustness (NLP literature). We then discuss limitations of the standard training paradigm in supervised machine learning and theorize about instruction-based models as a solution to increase validity. Instruction-based models are language models that receive explicit instructions for their task verbalised in plain text in addition to the fine-tuning data. We theorize that an explicit verbalisation of the task and the concept of interest helps a model learn the task more robustly and reduces the model's reliance on group-specific language patterns from the fine-tuning data (section 2).

We then test our assumptions empirically, by testing the interaction between bias and test-set performance of different models. We fine-tune three types of text classifiers on texts from four datasets and nine different groups under different conditions, resulting in 312 different fine-tuning runs.[1] Our results show that all types of models are susceptible to learning group-specific language patterns and that fine-tuning on biased data (from one group member) reduces performance on representative test sets (from all group members). On average, however, these effects are surprisingly small. In particular, we show that models receiving instructions as an additional input are particularly robust against biases from the fine-tuning data and are best at producing valid measurements across different groups. A language model without instructions (BERT-base) sees its average test-set performance decrease by 1.7% F1 macro when trained on biased data compared to random data. Its probability of making an error on group members it has not seen during training increases from 26% to 31.1%. An instruction-based model (BERT-NLI) sees its performance drop by only 0.4% F1 macro when trained on biased data and its probability of making an error on group members it has not seen during training increases only from 18.5% to 19.3%.

[1] The full reproduction code is available at https://anonymous.4open.science/r/language-model-bias-validity-D6E4/

2. Measurement validity and bias in computational text analyses

2.1 Measurement validity and supervised machine learning

Validity is a notoriously ambiguous term. Adcock and Collier "have found 37 different adjectives that have been attached to the noun 'validity' " (2001, p. 530).[2] For the purpose of this paper, we use the term "measurement validity" as the main type of validity based on Adcock and Collier (2001). Adcock and Collier's conceptualization is broadly applicable to qualitative and quantitative research, and we show that it also provides an excellent organizing framework for computational text analysis methods. Informally speaking, a measurement is valid, when it measures what the researcher wants it to measure (Adcock & Collier, 2001). To systematize this informal definition, it is helpful to make the process of measurement explicit using the example of a supervised machine learning project. SML projects in the social sciences normally start with a substantive research interest that requires the measurement of a background concept. The background concept could be 'populism'. The substantive research question could be whether 'populism' increased across time and countries during the COVID-19 pandemic. As a first step, researchers then need to narrow down the background concept for which many different definitions exist (level 1 in figure 1), into a systematized concept with a clear definition (level 2). They could, for example, follow an ideational definition of 'populism' where politics is seen as a struggle between the virtuous and homogeneous people and the selfish and corrupt elites (Cocco & Monechi, 2021, p. 3). The researchers then need to operationalize this concept to create a quantitative indicator (*a measurement*) of 'populism' (level 3). They could decide to measure expressions of populist ideas in texts and operationalize it by counting the occurrence of populist

[2] This is maybe unsurprising as the term is widely used by different disciplines and has unavoidably become part of contestations of how to conduct scientific inquiry. The term was coined in the 1950s by the psychology literature with the typology of content, criterion and construct validity and later converged to a unitarian definition of construct validity as the overarching term (Adcock 2001: 536-537). The causal inference literature emphasizes the terms internal and external validity (Cook and Campbell 1979, based on Adcock 2001: 529). The content analysis literature discusses at least 12 different types of validity (Krippendorff, 2018). The political science text-as-data literature (the focus of this paper) uses yet another set of complementary, but also different terms (Benoit, 2020; Grimmer et al., 2022; Grimmer & Stewart, 2013). Only very few interdisciplinary efforts for harmonizing terminologies from a computational perspective exist (Jacobs & Wallach, 2021).

sentences in party manifestos. To implement this operationalization, they would create a codebook with instructions for identifying "populist" vs. "non-populist" language. Based on this codebook, research assistants would then annotate ('score') several hundred sentences based on the pre-defined classification scheme (level 4). If the textual corpus is too large for full manual annotation, the researchers could then train a supervised classifier to automatically classify ('score') the remaining sentence in the full corpus of thousands of party manifestos. The researchers then need to aggregate the classifier's predictions ('scores') for individual sentence into the indicator, for example by simply summing up the number of populist sentences per year and per country (level 4 to 3). The resulting indicator (*the measurement*) then enables statements such as 'in year Y parties from country A used more populist language than in year Z or than in country B'. If everything went well, this indicator provides a valid measurement of populism – that is: an increase in the indicator indicates a real increase in populism (as defined by the researcher) in a given country or year.

Figure 1: Main steps for creating a valid measurement

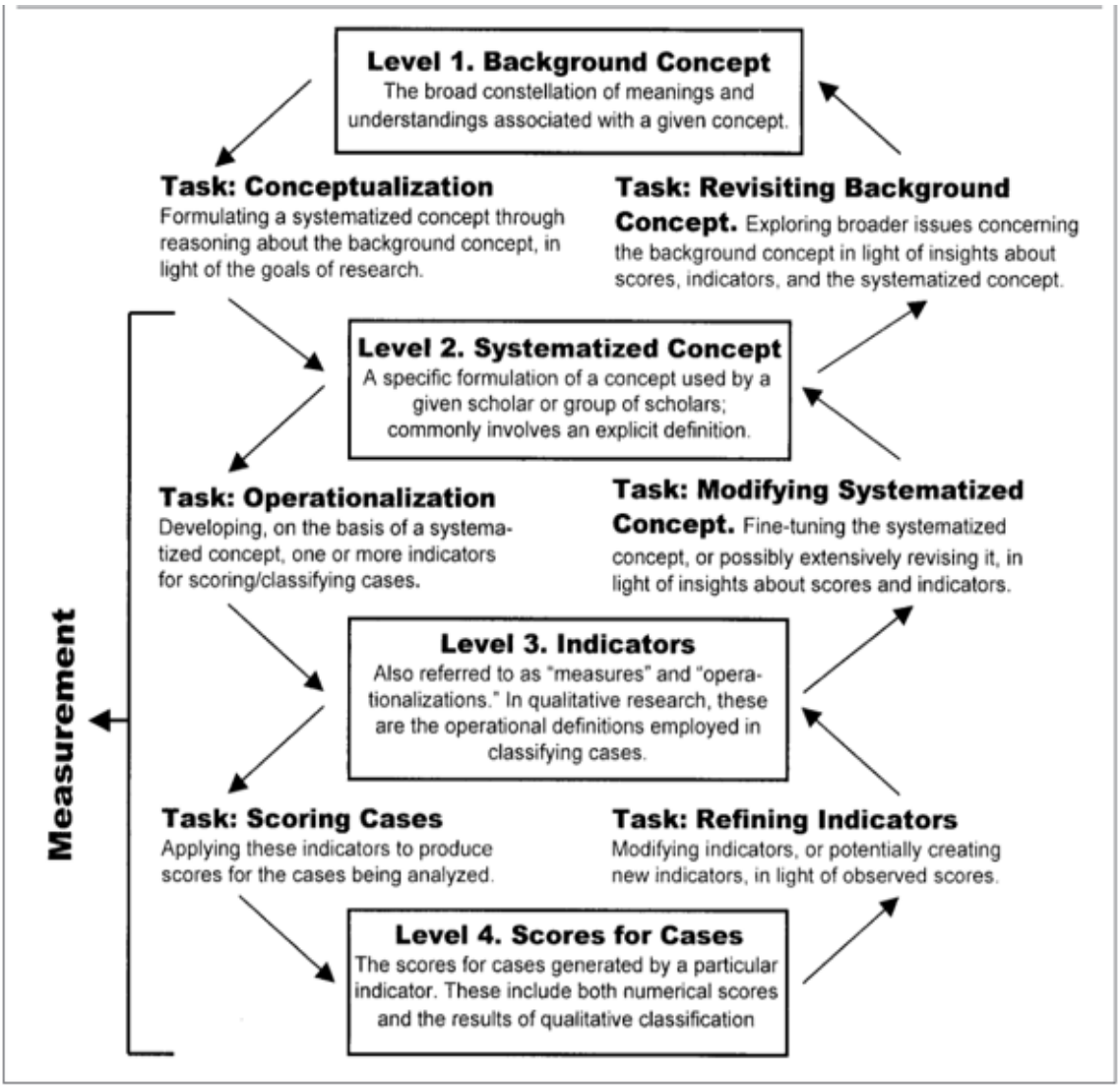


Figure from Adcock et al. 2001, p. 531

It is probably obvious to the reader that many things can go wrong during this process. Problems at any level can impact validity and therefore skew substantive conclusions. The key objective of validation is to ensure that this does not happen.[3] As there is a broad literature on measurement validity, this paper only focusses on two specific aspects linked to measurements derived from supervised machine learning: First, we investigate how group-specific biases in the training data can

[3] Note that we follow Adcock et al. (2001) in only using one overarching term for 'validity' (measurement validity), while there are different procedures of 'validation', which help establish measurement validity. The text-as-data literature uses roughly four validation procedures: Content validation, test-set validation, hypothesis validation and correlation validation.

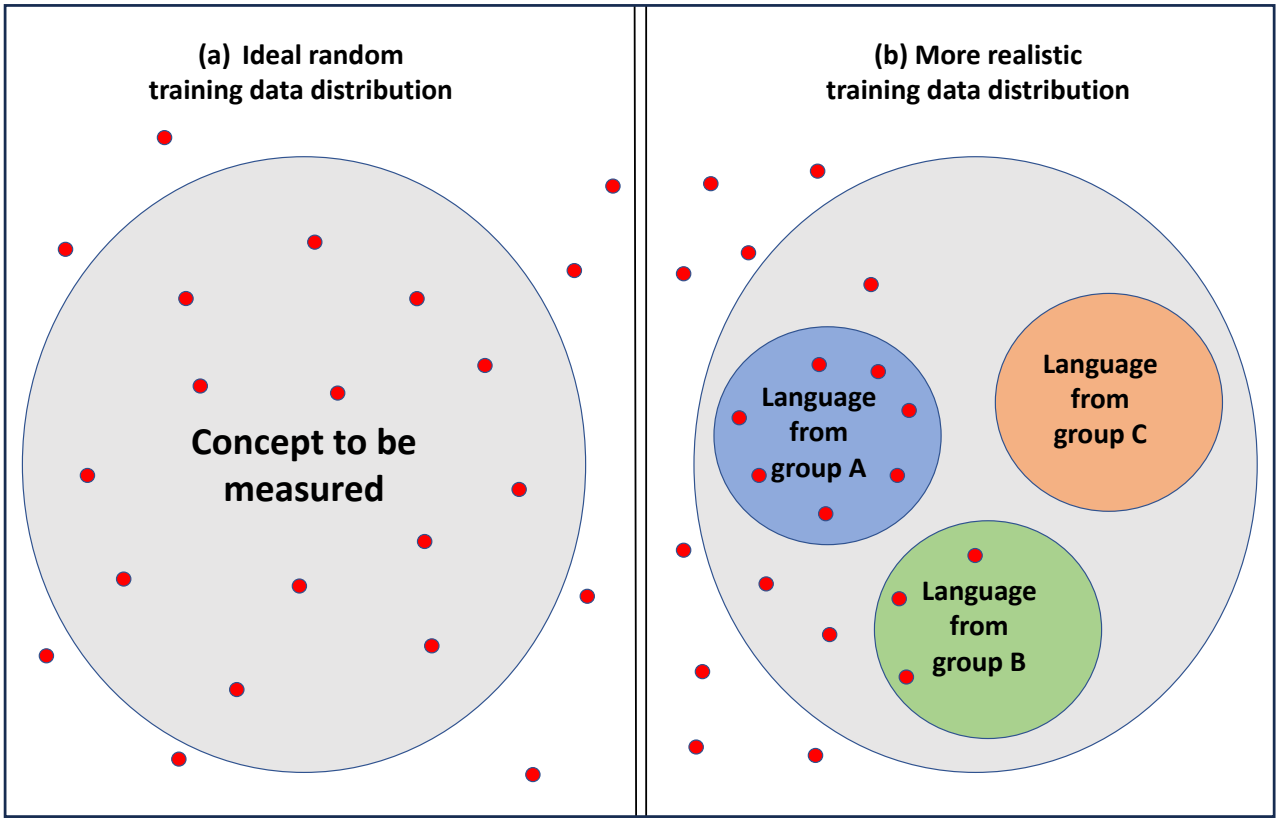
impact the scoring error at level 4. Second, we hypothesize that instruction-based models can create a systematic link between the scoring process (level 4) and the systematized concept (level 2), further reducing measurement error and increasing measurement validity. As the main type of validation in this paper, we use test-set validation combined with additional statistical tests.

2.2. Robustness against group-specific patterns as a precondition for validity

Why are group-specific patterns relevant for validity? For a standard machine learner, the only source of information for learning the concepts of interest and the related scoring task is the training data. A classical logistic regression does not have any prior knowledge about language, our task or concepts. Everything it learns about our concept of interest comes exclusively from the training data. This is similar for models like BERT (Devlin et al., 2019). While BERT models have language knowledge from their pre-training phase, a large part of the information about our task comes from the fine-tuning data. These models are designed to find *any* patterns in the training data that help them reduce their error on their training data. They are "stochastic parrots". A standard logistic regression only has a chance at fully learning the concept of interest if the training data comprehensively represents all ways of expressing the concept of interest. This ideal scenario is quite unrealistic (figure XXX, left). A more realistic scenario is illustrated in the right part of figure X. In practice, researchers will have access to imbalanced data from a few social groups (e.g. countries or milieus) and the machine will try to learn the concept of interest from the (group-specific) patterns in this data. This can work well, if the concept should only be measured in these specific groups, but it will work less well on data from groups outside of the training data that express the same concept differently. The more concepts (classes) a classifier is supposed to identify, and the more substantively relevant groups are in the data, the higher is the risk of the model learning (spurious) group-specific patterns instead of the

actual concepts we want to measure.[4] This directly impacts validity: measurements are only valid if they measure what we think they measure.

Figure 2: Semantic space and possible training data distributions



Left: The ideal training data distribution covers all main ways of expressing the concept of interest. Right: A more realistic distribution where several (group-specific) ways of expressing the concept of interest are missing. The grey area represents the entire semantic space for expressing the concept. Red dots represent individual training data points.

There is a broad literature on how machine learning models rely on simple patterns (shortcuts) in their (pre)training data to solve their tasks, instead of truly understanding the underlying task and concept (Du et al., 2022). Language models like BERT one-sidedly base their predictions on specific keywords ('lexical bias'), the positioning of words e.g. if predictive words always occur in the beginning of the text ('position bias'), overlapping terms for bi-text tasks ('overlap bias'), or specific writing styles that

[4] Note that figure 2 only illustrates a binary classification task for identifying if a text contains one concept or not. In practice, most classification tasks comprise multiple classes, exacerbating the risk of biased data distributions.

are irrelevant for the semantics of the task of interest ('style bias'). These biases remain hidden during test-set validation, if the test data is sampled from the same distribution as the (biased) training data. A model can successfully rely on these shortcuts, as long as they also work in the data to be analysed. In many practical contexts, however, the performance of these models will be reduced, if the data to be analysed comes from a different data distribution. Patterns learned from training data from one type of text might not work on other types of texts. These issues are also often discussed under the terms "robustness" against spurious patterns or "generalisation" beyond the training data (Wang et al., 2022).

Situations where these shortcuts and spurious patterns are linked to specific social groups are analysed in the fairness literature (Caton & Haas, 2020; Mehrabi et al., 2021; Pessach & Shmueli, 2022). A common view is that a model is fair (unbiased), if it performs equally across different social groups and unfair (biased) if it performs worse on specific groups (such as different countries). A classifier can, for example, perform worse on texts in specific languages or countries. The fairness literature proposes several different methods for identifying and remedying biases, from disaggregating metrics by groups, to different data pre-processing or post-processing techniques (Caton & Haas, 2020). One key dilemma is that methods for increasing fairness tend to increase measurement error (Caton & Haas, 2020, p. 18) (Caton & Haas, 2020, p. 18) and therefore impact measurement validity. There are some social science papers investigating the issue of spurious patterns for supervised machine learning (Hirst et al., 2014; Jankowski & Huber, 2023), but viable solutions are still lacking in our toolkit.

2.3 Increasing measurement validity with instructions

How can the issue of learning biased patterns be addressed? We argue that the standard fine-tuning procedure is at the core of these validity issues. As our argument applies both to training classical models as well as fine-tuning BERT models, we use the terms 'fine-tuning' and 'training' synonymously

for simplicity. From a validity perspective, the information the model receives during fine-tuning is incomplete. The only two types of inputs the models receive during fine-tuning are (1) N example texts and (2) a meaningless numeric label attached to each text representing one of K classes. By pure design of this procedure, the model is then forced to search for any patterns in the example texts that allow it to separate the texts into these K classes.

This is effectively equivalent to the following scenario: we pick a random person from the street and lock them into a room with 1000 texts with one of the numbers 1, 2, 3 or 4 written on top of each text. The only note left for the person is "Here are 1000 texts which I have split into 4 categories for you. Please learn the meaning of these categories. I will later ask you to identify them in other texts". We then unlock the door after a few hours, and quiz them if they have understood the complex social constructs we want to measure. Would we be surprised if they did not understand that "1" clearly represents our favourite complex social construct (e.g. populism) that needs to be carefully distinguished from the construct "2", which we had never told them anything about? While the obvious answer is "no", this is essentially how most supervised machine learning works. In the case of a logistic regression, the random person is a simple equation that has no prior knowledge of language or the task we are interested in (an 'illiterate and ill-instructed parrot'). In the case of BERT, it is a more complex set of equations and matrices that represent language, but without prior knowledge of the task and concept we are interested in (a 'literate but ill-instructed parrot'). It seems unsurprising that these systems can produce invalid measurements.

While this is how tasks are mostly taught to machine learners, there a very important difference when researchers teach tasks to crowd workers or research assistants. When teaching a task to human annotators, the main tool is a codebook with clear definitions and just few examples for the concept of interest and task. Throughout the annotation process, a diligent research assistant can refer to the codebook and anchor their annotation decisions in the explicit definitions of complex constructs.

The instruction paradigm from the NLP literature provides a way to implement a similar process with language models (Lou et al., 2023).[5] It follows almost the same steps as the pre-train-fine-tune paradigm, only that models are fine-tuned with task instructions as an additional, third input. Several different variants of this approach exist, from instructing GPT models (Brown et al., 2020; OpenAI, 2023), to prompting Masked Language Models like BERT (Schick & Schütze, 2021), to combining general tasks like Natural Language Inference (NLI) with BERT models (BERT-NLI) (Yin et al., 2019; Laurer et al., 2023a). In the NLP literature, these methods have mostly been discussed from the perspective of 0-shot or few-shot learning and only a few papers have investigated the robustness benefits of instruction-based models (Raman et al., 2023). Only very few papers have applied instruction-based models in the social sciences (Argyle et al., 2023; Laurer et al., 2023a, 2023b). We are not aware of a single paper investigating a systematic link to measurement validity.

From a validity perspective, the interesting feature of instruction-based models is that definitions of systematized concepts can be directly provided to the model as instructions. In practice, this means that the model is always fed a third input in addition to the standard two: (1) the text we want to analyze, (2) the desired output (e.g. a class label) and (3) instructions written in plain language, such as "Does this text contain populist language, describing 'the people' as virtuous and homogeneous or 'the elites' as selfish or corrupt?".

We theorize that this can provide a direct means for increasing the validity of supervised machine learning by directly linking level 2 with level 4 (figure 1). In this paper, we analyse the specific implications of instruction-based models for measurement validity and their robustness against bias from group-specific patterns. More specifically, we hypothesize that the instructions provide additional meaningful information to the model, enabling it to better learn a new concept of interest while relying less on patterns from the fine-tuning data. This hypothesis is, however, controversial. Evidence from the NLP literature does indicate that instruction-based models are more robust against

[5] A strand of literature uses the word "prompts" instead of "instructions".

spurious patterns from fine-tuning data, but some argue that this is linked to specific algorithmic properties of instruction-based models instead of the semantics of instructions (Raman et al., 2023; Webson & Pavlick, 2022). Fine-tuning a standard BERT-base on a new task entails deleting the task-specific head of the model and randomly reinitialising a new task head for the new classification task, while instruction-based models re-use all their parameters for new tasks and do not need to reinitialize parameters. Raman et al. argue that it is this algorithmic difference that makes instruction-based more robust against spurious patterns in training data instead of the semantics of instructions. We test both possible explanations empirically below.

3. Study design

We conduct our experiments on four datasets and nine types of groups (see table 1). Criteria for choosing datasets were: relevance for social science research; different types of tasks and concepts; texts from a diverse set of domains; availability of metadata for splitting the data in different social groups; sufficient data for training and testing across multiple groups.

Table 1: Datasets used in experiments

Dataset	Task & Concepts	Text domain	Groups	Data size
PImPo (Zobel & Lehmann, 2018)	Identify stances towards Immigration/Integration (4 classes: supportive, skeptical, neutral or not about immigration/ integration)	Party manifestos	10 party families, 14 countries, 3 decades	Train: 87168 Test: 6792
CoronaNet (Cheng et al., 2020)	Identify four types of policy measures against COVID-19 ('Public Awareness Measures', 'Restriction and Regulation of Businesses', 'Restrictions of Mass Gatherings', 'Health Resources')	Texts written by research assistants and copied from news sources	197 countries, 6 continents, 3 years	Train: 15326 Test: 3832
CAP-SotU (Policy Agendas Project, 2015)	Identify five topics ('Macroeconomics', 'Government Operations', 'Defense', 'International Affairs', 'Health')	US presidential speeches	2 phases (pre/post 1991), 2 parties (democrats/ republicans)	Train: 9248 Test: 2313
CAP-2 (SotU+Court) (Policy Agendas Project, 2014)	Identify five topics ('Domestic Commerce', 'Law and Crime', 'Civil Rights', 'Labor', 'Government Operations')	US presidential speeches & US court rulings	2 domains (speech / legal)	Train: 7708 Test: 1928

We compare the following classification models:

- Logistic regression as a representative for classical machine learning approaches ('illiterate and ill-instructed parrot');
- BERT-base[6] as a representative of standard transfer learning approaches ('literate but ill-instructed parrot'); and
- BERT-NLI as a representative for instruction-based approaches ('literate and instructed parrot'). For BERT-NLI, all experiments are run twice: once with meaningful and once with meaningless instructions. We call the second version BERT-NLI-void for short, as it receives instructions that are void of meaning. See the exact instructions used in appendix B.[7]

For all these datasets and classifiers, our experiments are then designed around our three main research questions.

1. *To what extent do group-specific language patterns (biases) impact measurement validity in supervised machine learning?*

We approximate the impact of group-specific biases in the training data in two steps: First, each classification model is trained on texts sampled from only one group member (e.g. only one country, "biased condition"). Second each model is also trained on texts randomly sampled across all group members ("random condition"). Classifiers from both conditions are then tested on the same fully random held-out test set that represents the dataset's real data distribution across all group members. We expect classifiers trained under the biased condition to perform less well on a representative test set, as these "biased" classifiers could only learn the concept of interest from the language of one group member, making it harder to extrapolate to other group members during

[6] We use Microsoft's DeBERTa-v3 model, a newer variant that strongly outperforms the original BERT model (He et al., 2021). We refer to it as "BERT" in the remainder of the text for simplicity. The same applies to BERT-NLI.
[7] The exact BERT-NLI model used is available at: [LINK ANONYMIZED]

testing. We conduct test-set validation with the standard classification metric F1 macro.[8] We call the difference in F1 macro between the biased and random condition for the same classifier the "bias penalty". This bias penalty indicates the loss in a classifier's ability to measure a concept of interest under biased conditions, i.e. when it only has access to language patterns from one group member during training.

This study design simulates extreme situations of bias in the training data by only sampling from one group member, while in practice researchers will often have access to data from more group members. This setup is designed to give us a clear idea of the impact of bias from group-specific language patterns. Section 4 will show that even in this setup the effect of these biases is relatively small. Another reason for this choice is to enable comparability across different datasets, as two of our datasets only have groups with maximum two group members (see table 1).[9] Moreover, this paper focusses on analysing the difference in robustness of different classifiers against biases across datasets and the role of instructions, instead of the reduction of bias in a specific case-study.[10]

Each model is always trained with 500 texts with balanced classes. Sampling with balanced classes is important, because prior research has shown that certain models perform better on imbalanced classes than others (Laurer et al., 2023a). As we want to compare which model is more robust against biases from group-specific patterns, we need to eliminate class imbalance as an intervening variable from the training data. A negative side-effect of this is that we cannot increase our training data above 500 texts. Some group members have very little data for some classes and with 500 texts there are still enough group members that have enough texts for minority classes.

---

[8] We use this metric because it gives equal weight to all classes. Class imbalance is an important issue in the social sciences and we assume that each class has the same substantive value independently of its size. See Laurer et al. (2023a) for an in-depth discussion of different classification metrics for social science use-cases.

[9] Only for the CoronaNet country group, we sample from three group members instead of one due to the low number of texts per country. The dataset contains very little data from individual (smaller) countries. Biasing the training data with three countries allowed us to introduce more biases from smaller countries.

[10] We also note that the effects of bias are already relatively small even if the training data comes from only one group. This is a central finding of this paper (see section 4). We had tested other experimental designs for analysing biases, such as introducing meaningless spurious tokens into texts. While these designs can show clear susceptibility to spurious language patterns and are used in the NLP literature, they are also less realistic. We therefore opted for analysing group-specific language patterns, which are more relevant for social scientists.

To reduce the influence of randomness, we repeat our training runs across 6 random seeds. In total, we train 3 types of models on 4 datasets, 9 types of groups, 2 degrees of bias (biased vs. random training data), across 6 random seeds. This leads to a total of 312 fine-tuned models and test-set results for testing the impact of group-specific biases. Note that we do not conduct hyperparameter searches for our experiments and use the recommended parameter values determined by an extensive hyperparameter search in Laurer et al. (2023a), as a hyperparameter search across this wide range of configurations (models, datasets, groups, bias) runs would be prohibitively expensive.

Note that test-set validation is only one procedure for ensuring measurement validity. Ensuring measurement validity is a complex multistep process that is specific to each measure and use-case (see section 2.1 and figure 1). For the purpose of our study across different datasets, we have to assume that these steps were well implemented for each dataset. Given this assumption, test-set validation with classification metrics is a good procedure that can be implemented comparatively across multiple datasets.[11]

2. *To what extent are different supervised machine learning approaches biased by group-specific language patterns in fine-tuning data?*

For our second analysis, we dive deeper into the bias of different classification models by analysing their classification predictions on the test data with a binomial mixed-effects regression. This analysis only uses the test results from the intentionally biased classifiers that were trained on data from one group member. We use the following variables: The (binary) dependent variable is the classification error, i.e. whether a given classification model made a mistake on a given test text nor not. The first (categorical) independent variable is the type of classifier, i.e. whether the classification prediction was made by a logistic regression, BERT-base or BERT-NLI. The second (binary) independent variable is whether the respective test text comes from the same group the classifier was trained on or not. If

[11] Other types of validation beyond test-set validation exist (e.g. content validation, hypothesis validation, correlation validation), but they cannot be properly implemented in a comparative study across datasets.

a row in our test data contains a prediction on a text from group member A by a classifier that was also trained on group member A, we flag it as a "biased row" in our data. Besides these two fixed effects, we also add a random effect to the binomial regression: the training run. We trained all types of classifiers multiple times across six random seeds for each group to account for randomness in classifier fine-tuning and therefore obtain multiple test results per classifier from six different training runs (see details below). To account for the non-independence of these observations and this hierarchical structure in our data, we include the identifier of the training run as a random effect in the mixed-effects regression.

The resulting binomial mixed-effects regression enables us to analyse the effect of the type of classifier and the effect of bias on error in the test data. The interaction between classifier type and biased rows and the resulting odds ratios provide indications of the degree of bias of different types classifiers.

3. Do meaningful instructions for language models reduce bias and increase measurement validity?

Lastly, we test the hypothesis that meaningful instructions provided to language models can function similarly to instructions provided to crowd workers, reducing group-specific biases and increasing measurement validity. For BERT-NLI, the researcher formulates instructions that verbalize each class.[12] For example, for the PImPo dataset, one instruction is "The text is sceptical of immigration/integration". To test our assumption, we need to test if it is actually the semantics of these instructions that improve our metrics, or other algorithmic properties of BERT-NLI. We therefore repeat all BERT-NLI training runs with meaningless instructions, such as "The text is about category A". See appendix B for all instructions used in our experiments.

[12] In the case of BERT-NLI, these instructions are normally called "hypotheses". See Laurer et al. (2023a) for more details.

4. Empirical results

RQ1: To what extent do these group-specific language patterns (biases) impact measurement validity?

For our first analysis, we conduct test-set validation with the standard classification metric F1 macro under a biased condition and a random condition (see red and blue dots in figures 3 and 4). The "bias penalty" is the difference in F1 macro between the biased and random condition, i.e. the distance between red and blue dots. The aggregated results in figure 3 show that, on average, the bias penalty is highest for the logistic regression classifier (2.3 percentage points) and shrinks from BERT-base (1.7%) to BERT-NLI (0.4%). The bias penalty for BERT-NLI is the smallest, indicating that its performance is least reliant on group-specific language patterns. Moreover, in line with previous research (Laurer et al. 2023a, 2023b), we find that BERT-NLI performs the best in terms of absolute test-set validation. BERT-NLI is best at learning the underlying concept of interest, while especially the logistic classifier fails to properly learn the classification tasks in the low data regime of 500 training texts. Note that we expect the difference between models to shrink as a higher quantity and diversity of texts is provided (Laurer et al. 2023a).

Figure 3: Test set validation and bias penalty in aggregate

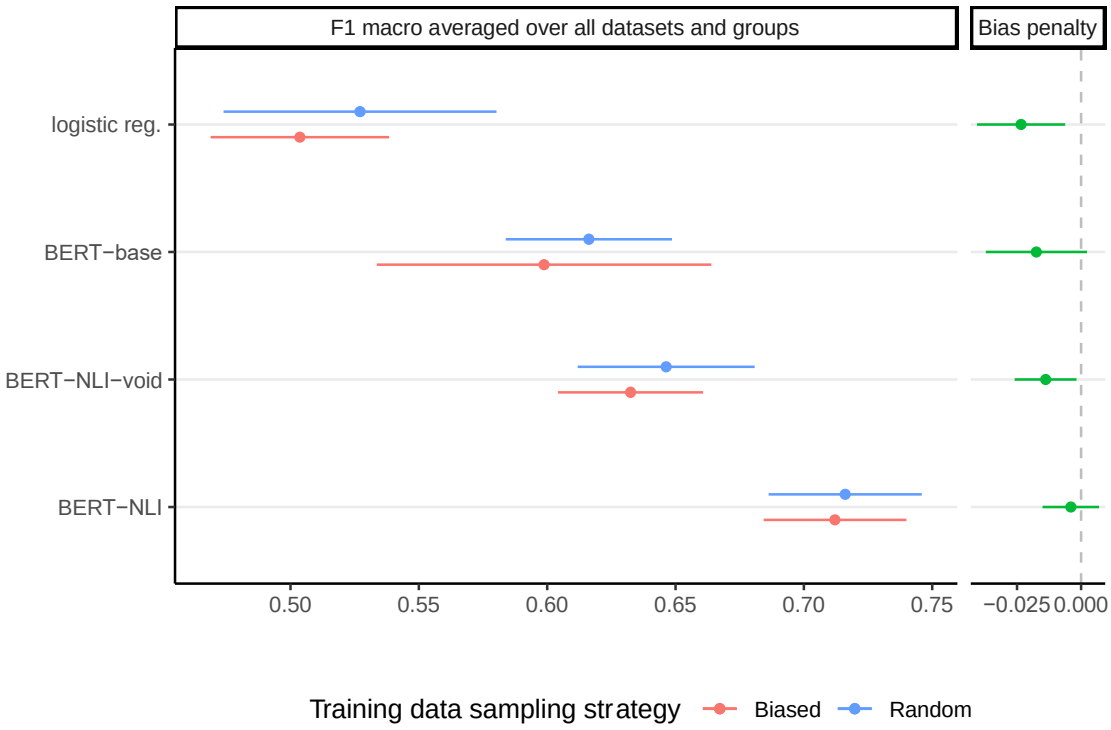
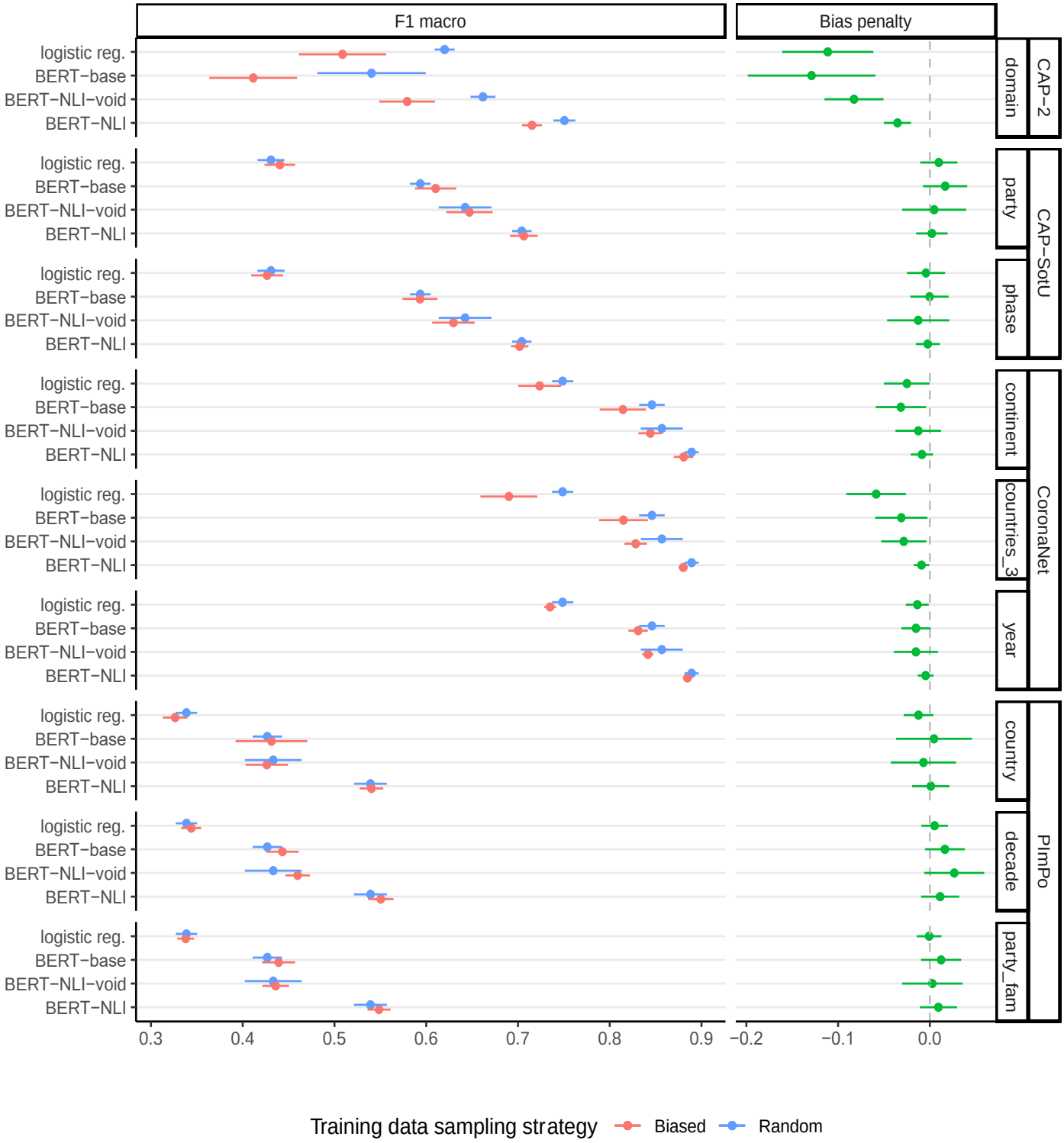


Figure 4 provides a more nuanced, disaggregated picture per group and dataset. The error bars display the standard deviation across six random runs with different random seeds and group members. For some datasets and groups, the bias penalty is very small while it is larger for others. We also notice that for a few combinations (PImPo and CAP-SotU party), some models perform on average better when trained on biased data. We assume that this is partly due to randomness and partly due to imbalance across group members in the test data. As the test data is randomly sampled from real-world datasets, it contains more data from some group members and little from others. The PImPo dataset, for example, contains more texts from certain party families because they talk more about migration and little data from others. For some group members with very little data, we could not sample 500 class balanced training texts and they were not included in the biased training runs. Depending on the dataset, the group members sampled for training can therefore also constitute the majority group members in the test data, which can explain that the performance of biased classifiers is sometimes higher than the classifiers trained on randomly sampled texts. We dive deeper into the issue of bias in the following analysis.

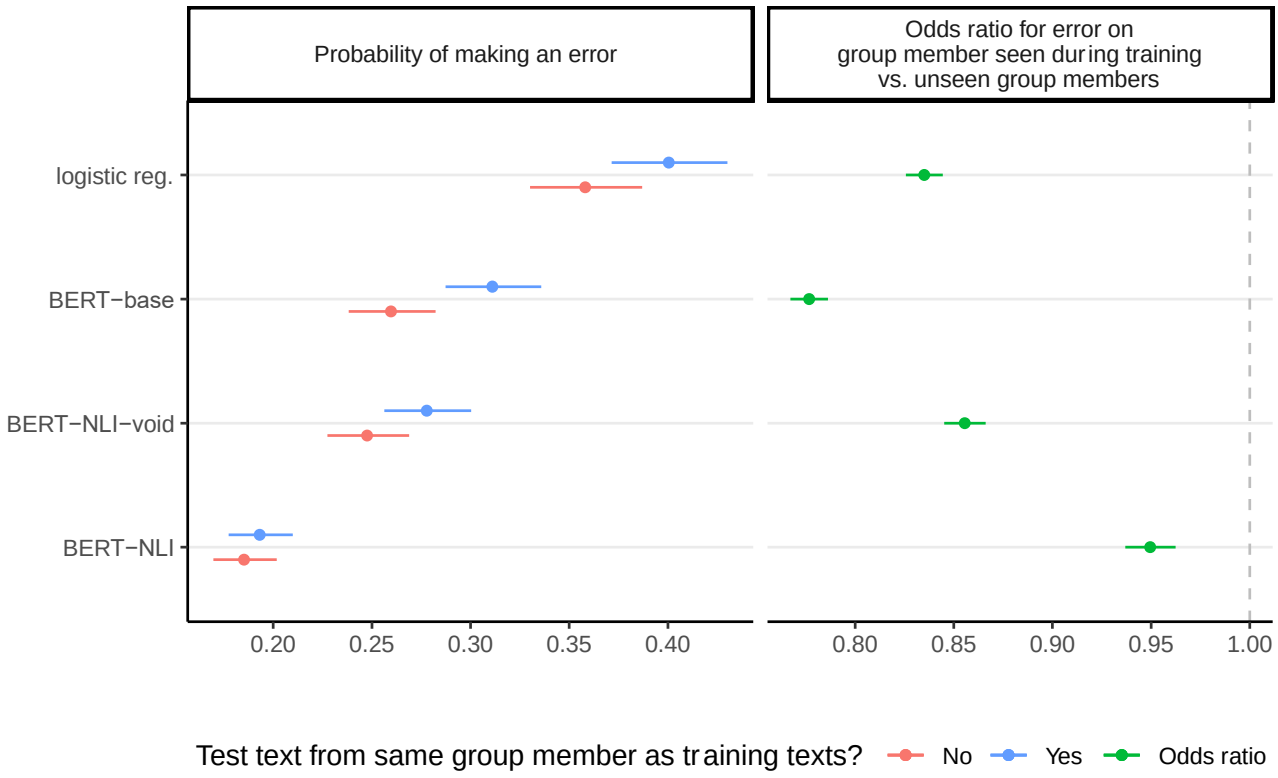
Figure 4: Test set validation and bias penalty by group



RQ2: To what extent are different supervised machine learning approaches biased by group-specific language patterns in fine-tuning data?

Figure 5 shows the results of the binomial mixed-effect regression described above. The regression analyses the predictions of classifiers trained on one group member (only intentionally biased classifiers). It shows how the type of classifier (independent variable) changes the odds of making an error (dependent variable) if a test text comes from the same (biased) group member the classifier has been trained on (interaction with second independent variable). The full regression tables and the figure disaggregated by group are available in appendix A. The red dots in figure 5 represent the classifiers' probability of making an error on test data that comes from the same group member as the classifier has been trained on, while the blue dot represents the error probability of error on data from group members the classifier has not seen during training.

Figure 5: Error analysis and bias



For the logistic classifier, we see that the odds of making an error decrease by a factor of 0.84 when the text comes from the same group member the classifier has seen during training. This is a clear indication of bias. While the probability of making an error on text from group members it has not yet seen during training is 40%, the probability of making an error on texts from group members it has seen during training is decreased to 35.8%. In accordance with our theoretical expectation from section 2, this bias effect is lowest for BERT-NLI. For BERT-NLI, the odds of making an error on biased texts only decrease by a factor of 0.95, i.e. from an error probability of 19.3% to 18.5%. Interestingly enough, the most biased model is BERT-base. BERT-base's odds of making an error are reduced the strongest by a factor of 0.78 on texts from group members it has already seen during training (26% probability of error) compared to group members it has not seen during training (31.1% probability of error). This indicates that BERT-base makes less mistakes than the logistic classifier overall, but a part of its performance advantage comes from learning more group-specific language patterns (overfitting). All results are highly significant.

To answer our second research question: all classifiers rely on group-specific language patterns to some extent. BERT-base overfits most strongly to these patterns. The logistic classifier relies slightly less on these patterns, but its lower ability to learn language patterns leads to the highest error rate overall. BERT-NLI is only marginally biased by group-specific language patterns and makes the least errors overall.

RQ3: Do meaningful instructions help reduce bias and increase measurement validity?

The results discussed above show that BERT-NLI is both less biased and performs better in terms of test set validity compared to a classical classifier and BERT-base. Why? One potential reason discussed in section 2 is that the instructions provided to the language model convey meaningful additional information and therefore reduce dependency on language patterns from the training data for learning a new task. An alternative explanation is that it is not about the meaning of instructions, but the fact that instruction-based models do not need to randomly reinitialize some parameters for new

tasks (see section 3). To test these different explanations, we now look at the results for BERT-NLI trained with meaningless instructions (BERT-NLI-void).

Based on the empirical results above, we conclude that both mechanisms contribute to BERT-NLI's improved performance. We find that BERT-NLI-void is slightly less biased than BERT-base in our regression analysis (figure 5) and it also performs better in terms of F1 macro with a smaller bias penalty (figure 3). BERT-base and BERT-NLI-void use the same underlying model and the only main difference are the training procedure with meaningless instructions. This indicates that the fact that BERT-NLI-void does not randomly re-initialise and re-learn task-specific parameters is an advantage over BERT-base. This effect is entirely unrelated to the meaning of instructions. At the same time, BERT-NLI-void is still forced to find any pattern in the training data to solve a task, because instructions like "This text is about category A" do not provide additional useful information. With meaningful instructions, BERT-NLI performs better both in terms of bias and overall F1 macro. As the only main difference between BERT-NLI and BERT-NLI-void is the meaning of the instructions, we conclude that the meaning of the instructions also contributes to its performance. Adding instructions such as "This text is positive towards immigration" make the model less dependent on learning patterns from the training data to understand new the task. Note that this does not mean that the model gains a deep understanding of the task like a research assistant. It does mean, however, that words like "positive" and "immigration" in the instruction enable the model to go beyond only patterns in the training data and leverage its internal representations of language to understand that the new task must have something to do with sentiment and migration.

This is also linked to another, more mechanistic difference between BERT-NLI and both BERT-base and BERT-NLI-void. Based on the hyperparameter search by Laurer et al. (2023a), we need to train BERT-base for many more epochs to achieve optimal performance compared to BERT-NLI. One epoch represents one iteration over the training data. Too many epochs lead to overfitting (the model relies too much on the patterns from the training data), while too few epochs mean that the model does not learn the new task properly. BERT-base and BERT-NLI-void need more time to find useful patterns

in the data that help them optimise for the new task and are therefore trained for more epochs. BERT-NLI has the instructions as a source of information and therefore needs less iterations over the training data to learn the new task. This is another reason why BERT-base (and BERT-NLI-void) overfit more easily to group-specific language patterns from the training data. We could train BERT-base for less epochs, which might reduce its biases, but would make it perform worse overall (see all hyperparameters in appendix C). BERT-NLI already learns the new task well with very few epochs, making it less prone to overfitting.

5. Limitations

Our analyses are subject to several limitations. First, we only analyse BERT-like Transformers (encoders) and no GPT-like generative Transformers (decoders or encoder-decoders), which have gained increasing popularity throughout 2023. GPT-like models are similar to BERT-NLI: they can also ingest instructions as a third input and they also re-use prior task knowledge from a universal task (text-token-prediction). We do not analyse GPT-like generative models in this paper for two main reasons. First, we are interested in creating measurements through text classification, while GPT-like models are designed for generative tasks and not specialised in classification. While any classification task can be reformulated as a generative task, generative models need to be much larger to obtain similar performance as a BERT-like model for text classification (Schick & Schütze, 2021; Xu et al., 2023). If we are only interested in creating measurements through classification, BERT-like classifiers are the suitable tool and we do not need the capability overhang from a generative model. Second, this size-requirement for good generators makes them less accessible and harder to handle on a hardware level. All BERT models used in this paper can be fine-tuned on a free GPU from Google Colab, as they are relatively “small” with around 214 million parameters. “Small” generators tend to have multiple billion parameters and can require multiple GPUs for fine-tuning. Despite these limitations,

we believe that generative models have great potential for social science applications, especially beyond text classification. We leave analyses of generative models to future work.

Second, we have analysed the problem of bias in fine-tuning data but did not analyse issues of bias in pre-training data. There is an established literature on bias in pre-trained models (Bender et al., 2021; Wang et al., 2022) as well as in NLI data (Gururangan et al., 2018) from the NLP fairness community. Our paper focuses on group-specific biases in the fine-tuning data that are relevant for comparative social scientists and measurement validity. Third, we used test-set validation as the main validation procedure. Several other types of validation exist that are useful for validation for specific case-studies (content validation, correlation validation, hypothesis validation), but are less suitable for validation and bias analyses across a wider set of datasets. As discussed in section 2, validation is a comprehensive process that needs to be adapted to each specific use-case. We focus on test-set validation as it is the gold standard procedure for validating supervised classifiers, it enables comparisons across datasets, and we complement this analysis with additional statistical tests.

6. Conclusion

This paper investigates the effect of group-specific biases in training data on measurement validity in supervised machine learning. We show that all types of classifiers learn group-specific biases. On average, the effects are however relatively small across 9 groups and 4 datasets with small and highly biased training sets. A classical logistic regression sees its F1 macro performance drop by 2.3 percentage points when trained on highly biased data instead of random data and its probability of making an error on group members it has not seen during training increases from 35.8% to 40% (0.84 odds ratio). BERT-base's test-set performance drops by 1.7% F1 macro when trained on biased data and its probability of making an error on group members it has not seen during training increases from 26% to 31.1% (0.78 odds ratio). BERT-NLI's performance drops by only 0.4% F1 macro when trained on biased data and its probability of making an error on group members it has not seen during training

increases only from 18.5% to 19.3% (0.95 odds ratio). We note that these effects are only averages and the bias effects are stronger for cases where language is very different between group members (especially for shifts from legal to speech language and partly between countries) and smaller for other groups (political parties or time).

We argue that the high level of robustness against bias and test-set validity of instruction-based models is due to two important characteristics. First, on an algorithmic level, instruction-based models do not need to delete and randomly re-initialize parameters, making them more robust. Second, they can ingest definitions of the task and concept of interest as plain text instructions, making them less dependent on (group-specific) language patterns in the training data and making it easier for them to learn the task and concept of interest.

Note that this paper only analyses advantages and limitations of different classifiers. When using supervised machine learning as a measurement tool for a specific substantive case-study, researchers should adhere to general good practices to ensure the validity of their measurements. Most of these good practices go well beyond the choice of classifier and could not be discussed in this paper. This starts with a proper definition of the concept of interest and task; to good training of annotators for creating high quality training data; to sampling representative and balanced train and test data; to aggregating classification predictions on individual texts into meaningful measurements.

As comparative researchers are often faced with situations where it is difficult to collect sufficient data for all relevant groups, we hope that our empirical analysis provides researchers with some optimism that, even when the training data is biased, instruction-based language models are good measurement tools. As language models improve and software and hardware become more accessible over the years, we believe that instruction-based language models will become an increasingly useful tool for social scientists to help them do their job: try and explain complex social phenomena with good measurements.

Bibliography

Adcock, R., & Collier, D. (2001). Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review*, *95*(3), 529–546. https://doi.org/10.1017/S0003055401003100

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 1–15. https://doi.org/10.1017/pan.2023.2

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Benoit, K. (2020). Text as Data: An Overview. In L. Curini & R. Franzese, *The SAGE Handbook of Research Methods in Political Science and International Relations* (pp. 461–497). SAGE Publications Ltd. https://doi.org/10.4135/9781526486387.n29

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Caton, S., & Haas, C. (2020). *Fairness in Machine Learning: A Survey* (arXiv:2010.04053). arXiv. https://doi.org/10.48550/arXiv.2010.04053

Cocco, J. D., & Monechi, B. (2021). How Populist are Parties? Measuring Degrees of Populism in Party Manifestos Using Supervised Machine Learning. *Political Analysis*, 1–17. https://doi.org/10.1017/pan.2021.29

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Du, M., He, F., Zou, N., Tao, D., & Hu, X. (2022). *Shortcut Learning of Large Language Models in Natural Language Understanding: A Survey* (arXiv:2208.11857). arXiv. http://arxiv.org/abs/2208.11857

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 107–112. https://doi.org/10.18653/v1/N18-2017

He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv:2111.09543 [Cs]*. http://arxiv.org/abs/2111.09543

Hirst, G., Riabinin, Y., Graham, J., Boizot-Roche, M., & Morris, C. (2014). Text to Ideology or Text to Party Status? In B. Kaal, I. Maks, & A. van Elfrinkhof (Eds.), *Discourse Approaches to Politics, Society and Culture* (Vol. 55, pp. 93–116). John Benjamins Publishing Company. https://doi.org/10.1075/dapsac.55.05hir

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385. https://doi.org/10.1145/3442188.3445901

Jankowski, M., & Huber, R. A. (2023). When Correlation Is Not Enough: Validating Populism Scores from Supervised Machine-Learning Models. *Political Analysis*, 1–15. https://doi.org/10.1017/pan.2022.32

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (Fourth Edition). SAGE.

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023a). Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, 1–33. https://doi.org/10.1017/pan.2023.20

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023b). Lowering the Language Barrier: Investigating Deep Transfer Learning and Machine Translation for Multilingual Analyses of Political Texts. *Computational Communication Research*, *5*(2), 1. https://doi.org/10.5117/CCR2023.2.7.LAUR

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 115:1-115:35. https://doi.org/10.1145/3457607

OpenAI. (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *55*(3), 51:1-51:44. https://doi.org/10.1145/3494672

Raman, M., Maini, P., Kolter, J. Z., Lipton, Z. C., & Pruthi, D. (2023). *Model-tuning Via Prompts Makes NLP Models Adversarially Robust* (arXiv:2303.07320). arXiv. http://arxiv.org/abs/2303.07320

Schick, T., & Schütze, H. (2021). It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *arXiv:2009.07118 [Cs]*. http://arxiv.org/abs/2009.07118

Wang, X., Wang, H., & Yang, D. (2022). *Measure and Improve Robustness in NLP Models: A Survey* (arXiv:2112.08313). arXiv. http://arxiv.org/abs/2112.08313

Webson, A., & Pavlick, E. (2022). Do Prompt-Based Models Really Understand the Meaning of Their Prompts? *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2300–2344. https://doi.org/10.18653/v1/2022.naacl-main.167

Xu, H., Lin, Z., Zhou, J., Zheng, Y., & Yang, Z. (2023). A Universal Discriminator for Zero-Shot Generalization. *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), 10559–10575. https://aclanthology.org/2023.acl-long.589

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. *arXiv:1909.00161 [Cs]*. http://arxiv.org/abs/1909.00161