Language Models as Measurement Tools

Laurer, Moritz

2024

DOI (link to publisher)
10.5463/thesis.742

document version
Publisher's PDF, also known as Version of record

VRIJE UNIVERSITEIT

Language Models as Measurement Tools

Using Instruction-Based Models to Increase
Validity, Robustness and Data Efficiency

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Sociale Wetenschappen
op woensdag 2 oktober 2024 om 9.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Moritz Laurer

geboren te Augsburg, Duitsland

promotor:	prof.dr. W.H. van Atteveldt
copromotoren:	dr. A. Casas Salleras dr. K. Welbers
promotiecommissie:	prof.dr T.B. Araujo dr. C. Baden prof.dr. M. Crosas prof.dr. P. Kerkhof dr. D. Nguyen prof.dr. D.C. Trilling

Summary

From millions of social media posts, to decades of legal text - more and more relevant information is hidden in digital text corpora that are too large for manual analyses. The key promise of machine learning is to automate parts of the manual analysis process. One popular method is supervised machine learning for text classification, where a model is trained on examples of manually categorized texts and learns to identify these categories in new texts. Computational social scientists have used this method to create measurements of concepts such as emotions, topics or stances at scale.

While measurement with supervised machine learning is established in the social science literature, there are important limitations that reduce the usefulness of established methods for many practical applications. First, these methods require large amounts of balanced training data to work well. Researchers, however, often only have limited resources for creating training data and need to tailor new data to each new research question. Second, older algorithms struggle with multilingual data. Researchers, however, need measurements that are equally valid for different cultures and languages. Third, they are susceptible to learning shortcuts and biased patterns from their training data, reducing the validity of measurements across social groups. Fourth, they can be difficult to use, making them only accessible to specialised researchers.

This thesis demonstrates how a recent innovation from the natural language processing literature can address these limitations: instruction-based language models. Chapter 2 shows how this type of model can reduce the required training data by a factor of ten compared to previous algorithms, while achieving the same level of performance across eight tasks. Chapter 3 demonstrates how these models require less than 2000 examples in two languages to create valid measurements across eight other languages and ten other countries. Chapter 4 shows how these models are more robust against group-specific biases. Their average test-set performance only decreases marginally when trained on biased data in experiments across nine groups from four datasets. Chapter 5 explains how these models can be universal classifiers that can learn any number of classification tasks simultaneously in tests across 33 datasets with 389 classes.

Acknowledgments

This PhD would have never come to fruition without the invaluable support of a wide range of people, only a few of whom fit on this page.

Firstly, I am grateful to my supervisors. When I reached out to professors based on their publications, I knew that personal fit would be at least as important as scientific fit. I was extremely lucky to have found both in my three supervisors. I am grateful to Wouter for agreeing to supervise me only based on an email and a call, and for his exceptionally pertinent feedback that probably saved me from months of chasing fruitless research ideas; to Andreu for his reliability in providing thoughtful feedback and always responding to my questions; to Kasper for his guidance on statistical methods and for being a pleasure to work with.

I am indebted to my partner Camille who not only supported me as my everyday companion, but also played an essential role in my shift to computational methods. Our countless discussions were instrumental in making my research more accessible and in finding practical usefulness in abstract scientific methods. Moreover, I am deeply indebted to my parents Andrea and Wolfgang, who, among many other things, made me know that I could always rely on them.

I am thankful to Andrea R. and CEPS, who supported this PhD through their flexibility and encouragement. I am also grateful for the financial support from the Heinrich Böll Stiftung, without whom the PhD could have taken much longer.

I also want to thank the members of the VU Amsterdam PolCom group for welcoming me and helping me navigate academic processes and conferences; Timo for our continuous discussions and his help in understanding academia; and Huy, Simon and Slava for welcoming me for my research stay at Hertie School.

Lastly, I would like to express my gratitude towards the open-source community. Without the countless people who (for fun, curiosity and status) provide free software and guidance online, this PhD would quite literally not have been possible. I have learned just as much from the authors of software libraries, documentation and forum posts as from academic papers.

Contents

Author Contribution Statement

Chapter 1
This chapter has not been submitted for publication elsewhere.

Moritz Laurer wrote the manuscript and is the sole author of this chapter. Atteveldt, Casas and Welbers acted as supervisors, providing valuable feedback on the text.

Chapter 2
A version of the chapter has been published as: *Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023). 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*

Moritz Laurer was the principal investigator of this study. Laurer wrote every version of the manuscript and the underlying code. Atteveldt, Casas and Welbers acted as supervisors, providing valuable feedback on different versions of the text.

Chapter 3
A version of the chapter has been published as: *Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023). 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*

Moritz Laurer was the principal investigator of this study. Laurer wrote every version of the manuscript and the underlying code. Atteveldt, Casas and Welbers acted as supervisors, providing valuable feedback on different versions of the text. Atteveldt also contributed code for visualisation.

Chapter 4
A version of the chapter has been published as: *Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2024). On Measurement Validity and Language Models: Increasing Validity and Decreasing Bias with Instructions. Communication Methods and Measures, 1–17. https://-doi.org/10.1080/19312458.2024.2378690*

Moritz Laurer was the principal investigator of this study. Laurer wrote every version of the manuscript and the underlying code. Atteveldt, Casas and Welbers acted as supervisors, providing valuable feedback on different versions of the text. Atteveldt also contributed code for visualisations and Welbers contributed code for statistical analyses.

Chapter 5
A version of the chapter has been published as a preprint as: *Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2023). Building Efficient Universal Classifiers with Natural Language Inference (arXiv:2312.17543). arXiv. https://doi.org/10.48550/arXiv.2312.17543*

Moritz Laurer was the principal investigator of this study. Laurer wrote every version of the manuscript and the underlying code. Atteveldt, Casas and Welbers acted as supervisors, providing valuable feedback on different versions of the text.

Chapter 6
This chapter has not been submitted for publication elsewhere.

Moritz Laurer wrote the manuscript and is the sole author of this chapter. Atteveldt, Casas and Welbers acted as supervisors, providing valuable feedback on the text.

Citing the entire thesis
The DOI of this thesis is http://doi.org/10.5463/thesis.742

Chapter 1

Introduction

"Training from a blank slate deprives our models of experience and the ability to interpret context. Ultimately, in order to come closer to the elusive goal of true natural language understanding, we need to equip our models with as much relevant knowledge and experience as possible."
(Ruder, 2019, p. 258-259)

Machine learning is marked by significant achievements and persistent challenges. Steady methodological progress has led to increasingly capable machines, exceeding expectations at incredible speed (Devlin, Chang, Lee, & Toutanova, 2019; LeCun, Bengio, & Hinton, 2015; OpenAI, 2023b). When it comes to applying machine learning to practical problems, however, researchers often struggle with issues of generalisation and validity. Success on machine learning benchmarks does not necessarily translate to success in practical applications in other disciplines (Bowman & Dahl, 2021; Kapoor et al., 2023; Liao, Taori, Raji, & Schmidt, 2021).

One such practical application is machine learning as a measurement tool in the computational social sciences. From a social science perspective, machine learning is not a purpose in itself, but only a tool for creating measurements. Machine learning is useful if it helps create valid measurements of concepts, such as emotions, topics, or stances. These measurements, in turn, can then inform explanatory models of complex social phenomena from elections to war (Benoit, 2020; Egami, Fong, Grimmer, Roberts, & Stewart, 2022; Grimmer & Stewart, 2013; Shah, Cappella, & Neuman, 2015; Theocharis & Jungherr, 2021; Van Atteveldt & Peng, 2021; Wallach, 2018).

A particularly relevant computational measurement tool is supervised

machine learning for text classification. In a supervised machine learning project, researchers start by defining concepts they are interested in measuring, for example four categories of emotions in social media posts. They then manually categorize example texts into these categories (classes) and train a classification model on these examples (training data). If this process is implemented well, the trained classification model learns to identify these classes in texts it has not seen before (test data). The classifier can then be used to predict the proportions of text in a large corpus that express these four emotions (the measurement). The resulting measurement can then be one variable in a broader explanatory model of a social phenomenon such as voter behaviour during elections (Grimmer, Roberts, & Stewart, 2021).

While measurement with supervised machine learning is established in the social science literature, there are important limitations that reduce the usefulness of established methods for many practical applications (Baden, Pipal, Schoonvelde, & van der Velden, 2022). First, these methods require large amounts of balanced training data to work well. Researchers, however, often only have limited resources for creating training data and need to tailor it to each new research question. Second, older algorithms struggle with multilingual data. Comparative researchers, however, need measurements that are equally valid for different countries and cultures. Third, they are susceptible to learning shortcuts and biased patterns from their training data, reducing the validity of measurements across social groups. Fourth, both older and newer models can be difficult to use in practice, making them only accessible to specialized researchers.

This thesis demonstrates how a recent innovation from the natural language processing literature can address these limitations: instruction-based language models. Chapter 2 shows how this type of model can reduce the required training data by a factor of ten compared to previous algorithms, while achieving the same level of performance across eight tasks. Chapter 3 demonstrates how these models require less than 2000 examples in two languages to create valid measurements across eight other languages and ten other countries. Chapter 4 shows how these models are more robust against group-specific biases. Their average test-set performance only decreases by 0.4% when trained on biased data in experiments across nine groups from four datasets. Chapter 5 explains how these models can be universal classifiers that can learn any number of classification tasks simultaneously in tests across 33 datasets

with 389 classes.

A key motivation for starting this thesis project was to explore innovations from machine learning research and to make them useful for addressing practical problems. In an attempt to make my research more broadly accessible, I have freely shared all my models and made them compatible with easy-to-use open-source libraries. On the day of submission of this thesis, my open-source models created during this endeavour have been downloaded more than 65 million times.[1]

1.1 Instruction-based models and transfer learning

The main innovation I investigate are instruction-based language models (Brown et al., 2020; Lou, Zhang, & Yin, 2023; Sanh et al., 2022; Wei et al., 2022). Instruction-based language models have two main characteristics.

First, they leverage the full spectrum of transfer learning by accumulating both prior *language* knowledge and *task* knowledge in their parameters (Ruder, 2019). Older classifiers like support vector machines or logistic regression start training without *any* prior knowledge. They need to learn the semantic difference between the words "war", "attack" and "tree" from their training data from scratch (no prior *language* knowledge). The only source of information for a new task is their training data (no prior *task* knowledge). Newer models like BERT gain *language* knowledge through pre-training (Devlin et al., 2019). They have an internal representation of the meaning of words but still need to learn any new task from their fine-tuning data from scratch (no useful *task* knowledge). Instruction-based language models, on the other hand, are (pre-)trained on a universal task such as text generation (Radford et al., 2019; Raffel et al., 2020) or Natural Language Inference, NLI (Dagan, Glickman, & Magnini, 2006; Yin, Hay, & Roth, 2019). These tasks are so general, that any specific task can be reformatted into the universal task. Being (pre-)trained on universal tasks enables instruction-based language models to learn new tasks without having to relearn task-specific parameters from scratch. They start with both useful *language and task* knowledge.

Second, these models can process instructions as an additional input. An instruction (or "prompt") is a natural language description of a task,

[1] 🤗 `https://huggingface.co/MoritzLaurer`

such as "Classify this text into one of two categories: positive or negative". Standard classifiers like BERT-base or support vector machines are not designed to process instructions and can only process two inputs: (1) The text to be analysed; (2) numeric labels for the desired output. By design, these standard models do not receive explicit information about their task and need to find any pattern in the input text that helps them correctly predict the numeric label. In contrast, instruction-based models can process instructions as a third input. The most prominent examples are probably the GPT models (Radford, Narasimhan, Salimans, & Sutskever, 2018). To fine-tune a GPT model for stance classification, for example, the inputs could be (1) a news article to be analysed, (2) a label like "positive" converted to numeric token IDs as the desired classification output, and (3) an instruction such as "Is this text positive or negative towards political elites?". This third input enables the model to learn new tasks better. Similar principles apply to other instruction-based language models like BERT-NLI, PET, or prompted RTD, which all process these three types of inputs in different ways (Lou et al., 2023; Schick & Schütze, 2021a; Xia, Artetxe, Du, Chen, & Stoyanov, 2022).

To understand the fundamental limitation of supervised machine learning without instructions, consider how difficult the standard supervised training procedure is even for humans. Imagine the following scenario: You want to measure eight types of emotions in a large corpus of a million social media posts. You recruit crowd workers to help you manually analyse these texts. To teach them your task, you send them an email with a few hundred example texts that you have categorized into your eight emotions of interest. In the email you send, the category for each text is only indicated by a number in the text's title. You do not provide any explanation of what the categories are about. Would the average crowd worker understand that you want to measure eight specific types of emotions based on a specific psychological theory only based on a few hundred texts and numbers?

While sending numbered texts without any explanations would obviously be a bad process for human learning, it is effectively how standard supervised machine learning works. Classification models traditionally only receive two inputs: example texts and meaningless numeric labels linked to each text, without any additional information about the meaning of the labels. The models are then trained to find any patterns in this training data that help it predict the correct numeric labels. Human annotators, by contrast, always receive a short codebook

with clear definitions and annotation instructions in addition to a few examples. In an attempt to model the human learning process more closely, instruction-based models are designed to ingest this third input: verbalized definitions of the task or classes of interest, i.e. instructions.

Several instruction-based models exist (Lou et al., 2023). This thesis focuses on one specific type: BERT-NLI (Yin et al., 2019; Yin, Rajani, Radev, Socher, & Xiong, 2020).[2] Its encoder-only architecture and training objectives are specialized for text classification, providing a good trade-off between model size and performance for measurement with text classification. While generative language models like T5 and GPT have capabilities beyond text classification, these capabilities are not necessary for classification (Schick & Schütze, 2021a; H. Xu, Lin, Zhou, Zheng, & Yang, 2023) and require infrastructure that is often beyond academic resources (at the time of writing). As many measurement tasks only require text classification, this thesis therefore only focuses on resource-efficient encoder-only models specialized in classification.

1.2 Dissertation overview

I develop my main arguments in four empirical chapters. Each chapter addresses one of the four key limitations of older algorithms outlined above and demonstrates how instruction-based models can help alleviate these limitations.

Chapter 2: How can we reduce data requirements for supervised text classification?

This chapter empirically demonstrates how the accumulation of both language and task knowledge in language models decreases the training data requirements for learning new social science tasks.

Early machine learning models were equations that had no prior knowledge at all (Pan & Yang, 2010; Ruder, 2019). Based on early work

[2]While there are many variants of BERT (the original BERT, RoBERTa, DistilBERT, AlBERT, DeBERTa etc.) their differences are less relevant for my research. They are all only slightly different encoder models that can be fine-tuned on downstream tasks. Their general design as encoders is the same (Yang et al., 2023). For simplicity, I use the term "BERT" to refer to such pre-trained encoder-only models. Any such model could be tuned to become a universal "BERT-NLI" model. The exact underlying models are specified in the respective chapters. The main base model used in this thesis is DeBERTaV3, an improved version of the original BERT (P. He, Gao, & Chen, 2021).

in the year 2000, word embeddings provided a way to create and store a form of language knowledge in vector representations (Bengio, Ducharme, & Vincent, 2000; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Since 2018, researchers started integrating language knowledge deeper into models like ULMFiT, ELMo or BERT (Devlin et al., 2019; Howard & Ruder, 2018; Peters et al., 2018). The main limitation of these models is, however, that they still need to learn any new task from scratch. Since then, researchers have developed techniques that also leverage a model's prior task knowledge, re-using all parameters from pre-training or pre-fine-tuning on a universal task like next-token-prediction or NLI (Radford et al., 2019; Raffel et al., 2020; Yin et al., 2019). While the advantages of newer models are relatively well understood for benchmark tasks in Natural Language Processing (NLP), this chapter demonstrates that these models can also address key challenges of social sciences tasks.

Social science tasks are characterized by challenges that arise when machine learning is only used as a (measurement) tool and is not the object of research itself. These tasks are characterized (a) by higher class imbalance reflecting the class distribution in real-world corpora; (b) by limited quantities of training data due to resource constraints as machine learning is only one aspect of a broader project; and (c) by types of classes that are informed by social scientist's explanatory research interests. This chapter compares test-set performance of four types of classification models across eight different social science tasks. Each model is trained on different training data samples increasing from 0 to 10000 texts from real-world, class-imbalanced corpora.

Empirical results show that the more prior knowledge a model uses, the better its performance on (imbalanced) data. Across these eight tasks, the BERT-NLI model fine-tuned on 100 to 2,500 texts performs on average 10.7 to 18.3 percentage points better than classical models that do not use transfer learning. As more data is provided, the performance difference decreases. The main practical conclusion for social science researchers is that newer models using more transfer learning and instructions can help save significant training data annotation resources during the measurement process.

Chapter 3: How can we derive valid measurement from multilingual texts?

This chapter investigates how multilingual language models and machine translation can be used to analyse multilingual texts.

Most computational text analysis methods are designed for English text (Baden et al., 2022). Comparative social scientists, however, often need to measure concepts across different languages and cultures. There are two main types of solutions to this problem. First, machine translation to English can be used to align all texts in the same language. This has long been the main solution used in the social sciences (de Vries, Schoonvelde, & Schumacher, 2018; Lucas et al., 2015). Second, newer multilingual transfer learning approaches create multilingual language knowledge in models, enabling them to ingest texts in multiple different languages simultaneously (Conneau et al., 2020; Conneau & Lample, 2019). Empirical evidence for the advantages and disadvantages of these different approaches in social science research is limited.

This chapter empirically compares machine translation to multilingual models, and combinations of both on two datasets with texts in 12 languages from 27 countries. In a first step, I compare test-set performance of four different types of models in both a high-resource and a low-resource multilingual setting. In a second step, I go beyond test-set validation and conduct hypothesis validation and correlation validation in a task measuring stances towards immigration across 10 political party families from 14 countries.

I find that the instruction-based BERT-NLI performs best among four model types, especially when little data is available. It performs best, both in terms of test-set and correlation validation. I do not find a very clear performance difference between multilingual models or English models combined with machine translation. I conclude that machine translation to English is probably the best approach for analysing multilingual texts, as it enables the use of newer models and makes manual validation easier for teams with limited language knowledge. Combining open-source machine translation with an English BERT-NLI model trained on 1674 texts from German and English can produce valid measurements on hundreds of thousands of texts from eight other languages and ten other countries. While these results are limited to two datasets, they indicate that future multilingual case studies can use machine translation to simplify the annotation process and could result in valid measurements with less than 2000 texts from a few countries.

Chapter 4: How robust are different models against group-specific biases?

This chapter investigates the robustness of different classification

models against group-specific biases in the training data, and theorizes about a systematic link between measurement validity and instruction-based models.

While (large) language models tout higher performance than classical 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 (Bender, Gebru, McMillan-Major, & Mitchell, 2021). 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 (Baden et al., 2022).

This chapter investigates this challenge in a comparative analysis across nine groups, four datasets and three types of models under biased and unbiased conditions. I train 312 text classifiers and analyse their robustness against group-specific biases and the validity of their outputs. I find that all types of models are susceptible to learning group-specific language patterns and that fine-tuning on biased data (from one group) reduces performance on representative test sets (from all groups). 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. Test-set validation and statistical bias tests indicate that they are best at producing valid measurements across different groups. The instruction-based BERT-NLI sees its average test-set performance drop by only 0.4% F1 macro when trained on biased data compared to random data. Its probability of making an error on groups it has not seen during training increases only by 0.8%.

Chapter 5: How can we build more efficient universal text classifiers?

The preceding chapters focus on making models better on a range of different individual tasks. This chapter investigates the universality of BERT-NLI, making it better at multiple tasks simultaneously and at new tasks it has not seen during training.

Generative large language models, perhaps most famously ChatGPT, have experienced an impressive boom in 2023 thanks to their ability to do many different text-related tasks with little to no task-specific fine-tuning (Chowdhery et al., 2022; OpenAI, 2023b; Touvron et al., 2023). These models pay their strong capabilities with resource and

infrastructure requirements beyond reach for academics. Moreover, using trillion parameter text generators for measurement projects that only require text classification is disproportionate.

This chapter demonstrates that smaller encoder-only models like BERT-NLI can be universal classifiers while being comparatively efficient and accessible. While only limited to text classification, they can learn any number of classification tasks simultaneously and generalize to unseen tasks without fine-tuning. This hands-on chapter guides the reader through several Jupyter notebooks from data preparation and cleaning, to training and evaluation of universal BERT-NLI models. The resulting model is trained on 33 datasets with 389 classes simultaneously. Its performance at doing new classification tasks without having seen training data (zeroshot classification) increases by 9.4% compared to NLI-only models. It can do inference on a laptop and the training can be reproduced in the browser for around 50 Euros on Google Colab.

Beyond individual chapters, the work on this thesis was always driven by the spirit of knowledge sharing and open-source (Van Atteveldt, Strycharz, Trilling, & Welbers, 2019). For all four empirical chapters, the full reproduction code is openly available on GitHub, along with materials from workshops and tutorials.[3] All models are freely shared online.[4]

[3] `https://github.com/MoritzLaurer`

[4] `https://huggingface.co/MoritzLaurer`

Chapter 2

Less Annotating, More Classifying

Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI

Abstract. *Supervised machine learning is an increasingly popular tool for analysing large political text corpora. The main disadvantage of supervised machine learning is the need for thousands of manually annotated training data points. This issue is particularly important in the social sciences where most new research questions require new training data for a new task tailored to the specific research question. This paper analyses how deep transfer learning can help address this challenge by accumulating 'prior knowledge' in language models. Models like BERT can learn statistical language patterns through pre-training ('language knowledge'), and reliance on task-specific data can be reduced by training on universal tasks like Natural Language Inference ('task knowledge'). We demonstrate the benefits of transfer learning on a wide range of eight tasks. Across these eight tasks, our BERT-NLI model fine-tuned on 100 to 2500 texts performs on average 10.7 to 18.3 percentage points better than classical models without transfer learning. Our study indicates that BERT-NLI fine-tuned on 500 texts achieves similar performance as classical models trained on around 5000 texts. Moreover, we show that transfer learning works particularly well on imbalanced data. We conclude by discussing limitations of transfer learning and by outlining new opportunities for political science research.*

Paper published as: *Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023). 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*

2.1 Introduction

From decades of political speeches to millions of social media posts - more and more politically relevant information is hidden in digital text corpora too large for manual analyses. The key promise of computational text analysis methods is to enable the analysis of these corpora by reducing the need for expensive manual labour. These methods help researchers extract meaningful information from texts through algorithmic support tools and have become increasingly popular in political science over the past decade (Benoit, 2020; Grimmer & Stewart, 2013; Lucas et al., 2015; Van Atteveldt, Trilling, & Calderon, 2022; Wilkerson & Casas, 2017).

Supervised machine learning is one such algorithmic support tool (Osnabrügge, Ash, & Morelli, 2021). Researchers manually create a set of examples for a specific task (training data) and then train a model to reproduce the task on unseen text. The main challenge of this approach is the creation of training data. Supervised models require relatively large amounts of training data to obtain good performance, making them a "nonstarter for many researchers and projects" (Wilkerson & Casas, 2017). Lack of data is particularly problematic in the social sciences where most new research questions entail a new task (task diversity) and some concepts of interest are only present in a small fraction of a corpus (data imbalance). Compared to the Natural Language Processing (NLP) literature, for example, political scientists are less interested in recurring benchmark tasks with rich and artificially balanced data. The ensuing data scarcity problem is probably an important reason for the greater popularity of unsupervised approaches in the social sciences. Unsupervised approaches are difficult to tailor to specific tasks and are harder to validate, but they do not require training data (Denny & Spirling, 2018; Miller, Linder, & Mebane, 2020).

This paper argues that this data scarcity problem of supervised machine learning can be mitigated through deep transfer learning. The main assumption of transfer learning is that machine learning models can learn 'language knowledge' and 'task knowledge' during a pre-training phase and store this 'knowledge' in their parameters (Pan & Yang, 2010; Ruder, 2019).[1] During a subsequent fine-tuning phase, they can then

[1]Note that we only use the word 'knowledge' to help create an intuitive understanding of transfer learning without too much jargon. Language models (i.e. pre-trained algorithms) do not 'know' or 'understand' anything in a deeper sense. The machine learning process is essentially a sequence of parameter updates to optimise the statistical solution of a very specific task. Some authors colloquially call

build upon this 'prior knowledge' to learn new tasks with less data. Put differently, a model's parameters can represent statistical patterns of word probabilities ('language knowledge'), link word correlations to specific classes ('task knowledge') and later reuse these parameter representations for new tasks ('knowledge transfer').

In the political science literature, the use of shallow 'language knowledge' through pre-trained word embeddings has become increasingly popular (Rodman, 2020; Rodriguez & Spirling, 2022), while the investigation of deep 'language knowledge' and models like BERT has only started very recently on selected tasks (Bestvater & Monroe, 2022; Licht, 2023; Widmann & Wich, 2022). We are not aware of political science literature on 'task knowledge'.

This paper therefore makes the following contributions. We systematically analyse: the benefits of transfer learning across a wide range of tasks and datasets relevant for political scientists; the importance of 'task knowledge' as a second component of transfer learning; the impact of transfer learning on imbalanced data; and how much training data, and therefore annotation labour, different algorithms require. Our insights can help future research projects estimate their data requirements with different methods.

To test the theoretical advantages of transfer learning, we systematically compare the performance of two classical supervised algorithms (Support Vector Machine, Logistic Regression) to two transfer learning models (BERT-base and BERT-NLI) on eight tasks from five widely used political science datasets.

Our analysis empirically demonstrates the benefits of transfer learning. BERT-NLI outperforms classical models by 10.7 to 18.3 percentage points (F1 Macro) on average when 100 to 2500 annotated data points are available. BERT-NLI achieves similar average F1 Macro performance with 500 data points as classical models with around 5000 data points. We also show that BERT-NLI performs better with very little training data $<\approx 1000$, while BERT-base is better when more data is available. Moreover, we find that 'shallow knowledge transfer' through word embeddings also improves classical models. Lastly, we show that transfer learning is particularly beneficial for imbalanced data. These benefits of transfer learning robustly apply across a wide range of datasets and tasks.

this internal parameter representation 'knowledge'. For a more formal discussion of transfer learning see Ruder (2019) Ruder (2019) and Pan and Yang (2010).

We conclude by discussing limitations of deep transfer learning and by outlining new opportunities for political science research. To simplify the re-use of BERT-NLI in future research projects, we open-source our code[2], general purpose BERT-NLI models[3] and provide advice for future research projects.

2.2 Supervised machine learning from a transfer learning perspective

2.2.1 Supervised Machine Learning in Political Science

The rich text-as-data literature demonstrates the wide variety of methods in the toolkit of political scientists: supervised or unsupervised ideological scaling; exploratory text classification with unsupervised machine learning; or text classification with prior categories with dictionaries or supervised machine learning (Benoit, 2020; Chatsiou & Mikhaylov, 2020; Grimmer & Stewart, 2013; Lucas et al., 2015; Van Atteveldt et al., 2022; Wilkerson & Casas, 2017). This paper focuses on one specific group of approaches: text classification with prior categories with supervised machine learning.

In the social sciences, supervised machine learning projects normally start with a substantive research question which requires the repetition of a specific classification task on a large textual corpus. Researchers might want to: explain Russian foreign policy by classifying thousands of statements from military and political elites into 'activist' vs. 'conservative' positions (Stewart & Zhukov, 2009); or understand delegation of power in the EU and classify legal provisions into categories of delegation (Anastasopoulos & Bertelli, 2020); or predict election results and need to classify thousands of tweets into sentiment categories to approximate twitter users' preferences towards key political candidates (Ceron, Curini, Iacus, & Porro, 2014). These research projects required the classification of thousands of texts in topical, sentiment or other conceptual categories (classes) tailored to a specific substantive research interest.

Using supervised machine learning to support this process roughly

[2]An easy-to-use Jupyter notebook for training your own BERT-NLI model and the full reproduction code is available at: `https://github.com/MoritzLaurer/less-annotating-with-bert-nli`

[3]Several models are available at `https://huggingface.co/MoritzLaurer`

involves the following steps: A tailored classification task is developed, for example through iterative discussions resulting in a codebook; experts or crowd workers implement the classification task by manually annotating a smaller set of texts (training and test data); a supervised machine learning model is trained and tested on this manually annotated data to reproduce the human annotation task; if the model's output obtains a desired level of accuracy and validity, it can be used to automatically reproduce the task on very large unseen text corpora. If implemented well, the aggregate statistics created through this automatic annotation can then help answer the substantive research question.

Political scientists have mostly used a set of classical supervised algorithms for this process, such as Support Vector Machines (SVM), Logistic Regression, Naïve Bayes etc. (Benoit, 2020). These classical algorithms are computationally efficient and obtain good performance if large amounts of annotated data are available (Terechshenko et al. 2020). Their input is usually a document-feature matrix which provides the weighted count of pre-processed words (features) per document in the training corpus. Solely based on this input, these models try to learn which feature (word) combinations are most strongly linked to a specific class (e.g. the topic "economy"). Several studies have shown the added value of these algorithms (for example Colleoni, Rozza, & Arvidsson, 2014; Osnabrügge et al., 2021; Peterson & Spirling, 2018).

The key disadvantage of these classical algorithms is that they start the training process without any prior 'knowledge' of language or tasks. Humans know that the words "attack" and "invasion" express similar meanings, or that the words "happy" and "not happy" tend to appear in different contexts. Humans also quickly understand the task "classify this text into the category 'positive' or 'negative'". Classical models on the other hand need to learn these language patterns and tasks from scratch with the training data as the only source of information. Before training, the SVM is only an equation that can draw lines into space. A SVM has no prior internal representation of the semantic distance between the words "attack", "war" and "tree". This lack of prior 'knowledge' of language and tasks is the main reason why classical supervised machine learning requires large amounts of training data.

A first solution to the 'language knowledge' limitation compatible with classical algorithms was popularised in 2013 with word embeddings (Mikolov et al., 2013). Word embeddings represent words that are often mentioned in similar contexts with similar vectors – a proxy for semantic

similarity. These embeddings can for example be used as input features for classifiers to provide them with a form of 'language knowledge' and have gained popularity in political science (Rodman, 2020; Rodriguez & Spirling, 2022). Word embeddings alone provide, however, only 'shallow language knowledge': first, the information they capture is limited. The vector of the word "capital" is the same, whether it appears next to the word "city", "investment" or "punishment". Second, the improvement, which word embeddings offer for classical algorithms is only a different input layer: word embeddings instead of e.g. TF-IDF as input. Newer models integrate word embeddings into stacked layers of many additional vectors (parameters). These multi-layered, 'deeper' architectures are designed to store more 'knowledge'.

2.2.2 Deep Transfer Learning

Deep transfer learning tries to create 'prior knowledge' by splitting the training procedure in roughly two phases: pre-training and fine-tuning (Howard & Ruder, 2018). First, an algorithm is pre-trained to learn some general purpose statistical 'knowledge' of language patterns in a wide variety of domains (e.g. news, books, blogs), creating a language model. Second, this pre-trained model is fine-tuned on annotated data to learn a very specific task.[4]

Transfer learning therefore has two important components (Pan & Yang, 2010; Ruder, 2019): (1) learning statistical patterns of language (language representations) and (2) learning a relevant task (task representations). Both types of representations are stored in the parameters of the model.

For learning general purpose language representations, the most prominent solution is BERT (Devlin et al., 2019) which is a type of Transformer model (Vaswani et al., 2017). Transformers like BERT are first pre-trained using a very simple task such as Masked Language Modelling (MLM), which does not require manual annotation. During MLM, some words are randomly hidden from the model and it is tasked with predicting the correct hidden words. The overall objective of this procedure is for the model's parameters to learn statistical patterns of language (language representations) such as semantic similarities of

[4]This describes the focus of the main steps. In practice, pre-training also involves learning (less relevant) task(s) and fine-tuning also involves learning the language of specific domain(s) (e.g. legal or social media texts).

words or context-dependent ambiguities from a wide variety of texts (see appendix B1 for details).

While sizeable performance increases with BERT-base models are possible based on its 'language knowledge' (Devlin et al., 2019), data requirements are still relatively high. Widmann and Wich (2022), for example, show strong performance gains for an emotion detection task, but point out that the amount of training data is still an important limitation and that classes with less data underperform. An important reason for this is that the pre-training task BERT-base has learned (MLM) is very dissimilar to the actual final classification tasks researchers are interested in. This is why the last, task-specific layer of BERT (the task head tuned for MLM) is normally deleted entirely and reinitialised randomly before fine-tuning – which constitutes an important loss of 'task knowledge' (see appendix B for details on BERT's layered structure). BERT then needs to be fine-tuned on manually annotated data, to learn a new, useful task and each of its classes from scratch.

2.2.3 BERT-NLI – Leveraging the Full Potential of Deep Transfer Learning

More recently, methods have been proposed which do not only use prior 'language knowledge', but also prior 'task knowledge' of Transformers.[5] There are several different approaches using these innovations (Brown et al., 2020; Raffel et al., 2020; Schick & Schütze, 2021a). This paper uses one approach, based on Natural Language Inference (NLI), first proposed by Yin et al. (2019) and later refined for example by S. Wang, Fang, Khabsa, Mao, and Ma (2021).

What is NLI? NLI is a task and data format, which consists of two input texts and three output classes. The input texts are a 'context' and a 'hypothesis'. The task is to determine if the hypothesis is True, False or Neutral given the context.[6] A hypothesis could be "The EU

[5]Note that the transfer of 'task knowledge' is not inherently limited to Transformers. Osnabrügge et al. (2021) show that the task learned by a Logistic Regression trained on the Manifesto Corpus can be applied to a different target corpus and that datasets with broadly useful tasks can be re-used with classical models. Transfer learning is not an 'either-or' category, but can be handled by different models to different extents.

[6]Note that there is some variation in how the input texts and classes are called in the literature. NLI can also be called Recognising Textual Entailment (RTE), the 'context' can be called 'premise' and the three classes can be called 'entailment', 'contradiction', 'neutral' (Dagan et al., 2006; Williams, Nangia, & Bowman, 2018).

Table 2.1: Examples of the NLI task

Hypothesis	**Context**	**Class**
The EU is trustworthy	The EU has betrayed its partners during the negotiations on Sunday	False
The EU is trustworthy	The US has betrayed its partners during the negotiations on Sunday	Neutral
The EU is trustworthy	Civil society praised the EU for reliably keeping its promises.	True

is trustworthy" with the context "The EU has betrayed its partners during the negotiations on Sunday". In this case, the correct class would be False, as the context contradicts the hypothesis. Note that it is not about finding the objective truth to a scientific hypothesis, but only about determining if the context string entails the hypothesis string. See table 1 below for examples.

NLI has three important characteristics from a transfer learning perspective: It is data-rich, it is a universal task, and it enables label verbalisation. First, NLI is a widely used and data-rich task in NLP. Many NLI datasets exist, and crowd-coders have created more than a million unique hypothesis-context pairs. Using this data, the pre-trained BERT-base can be further fine-tuned on the NLI classification task, creating BERT-NLI. Our BERT-NLI models are trained on a concatenation of eight general-purpose NLI datasets (around 1.2 million texts) from the NLP literature (see appendix B3 for details).

Second, NLI is a universal task. Almost any classification task can be converted into an NLI task. Take the text "We need to raise tariffs" and our task could be to classify this text into the eight topical classes of the Manifesto Corpus ("Economy", "Democracy", …). BERT-NLI can always only execute the NLI task: predicting one of the classes True/False/Neutral given a context-hypothesis pair. We can, however, translate the topic classification task into an NLI task by expressing each topical class as a 'class-hypothesis', e.g. "It is about economy", "It is about democracy" etc. We can then take "We need to raise tariffs" as context and test each of the class-hypotheses against this context. Each context-hypothesis pair is provided as input to BERT-NLI, which predicts the three NLI classes True/False/Neutral for each class-

We use the simplified vocabulary based on the instructions shown to crowd workers.

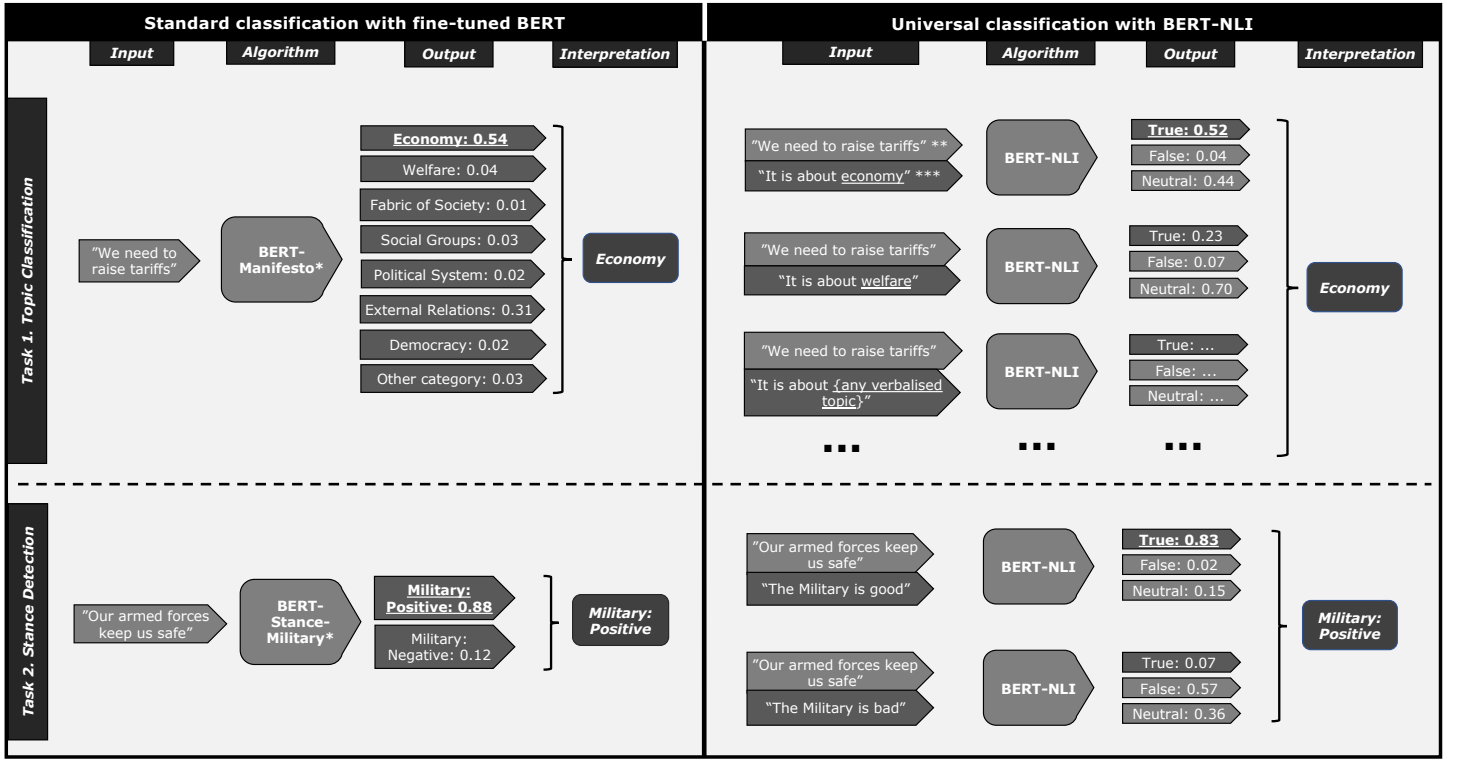
hypothesis. We then select the topical class via the class-hypothesis that BERT-NLI predicts to be the 'truest'. Note that when we repurpose BERT-NLI for other tasks like topic classification, the class-hypotheses do not have to be actually 'true' in a deeper sense. The objective of reusing the classes of BERT-NLI for other tasks is only to identify the most likely downstream class relevant for the new task. The predictions for the NLI classes False and Neutral class are ignored. Figure 1 illustrates how this approach enables us to solve almost any classification task with BERT-NLI.

Using a universal task for classification is an important advantage in situations of data scarcity. Both classical algorithms and BERT-base models need to learn the target task the researcher is interested in from scratch, with the training data as the only source of task-information. They can then only solve this very specific task. With the universal BERT-NLI classifier, almost any task can be translated into the universal NLI task format. BERT-NLI can then fully re-use the 'task knowledge' it has already learned from hundreds of thousands of general-purpose NLI context-hypothesis pairs. No task-specific parameters need to be randomly reinitialised in the task head. No 'task knowledge' is lost.

This is also linked to the third important characteristic of NLI classification: label verbalization (Schick & Schütze, 2021a). Remember that human annotators always receive explicit explanations of each class in form of a codebook and can use their prior knowledge to understand the task without any examples. Standard classifiers, on the other hand, only receive examples linked to an initially meaningless number for the respective class (both classical algorithms and BERT-base). They never see the description of the classes in plain language and need to statistically guess what the underlying classification task is, only based on the training data. With the NLI task format, the class can be explicitly verbalised in the hypothesis based on the codebook (see figure 1). More closely imitating human annotators, BERT-NLI can therefore build upon its prior language representations to understand the meaning of each class more quickly. Expressing each class in plain language provides an additional important signal to the model.

As we will show in section 3, the combination of Transformers, self-supervised pretraining, intermediate training on the data-rich NLI task, reformatting of target tasks into the universal NLI task and label verbalisation can substantially reduce the need for task-specific training data.

Figure 2.1: Illustration of standard classification vs. universal NLI classification



* BERT models fine-tuned on a specific dataset respectively. They can only predict the exact classes they have learned from their respective training data.
E.g. topics from the Manifesto corpus, or stances towards the military.

** Context: Any input text treated as context for the class-hypothesis.

*** Class-hypothesis: Class verbalised by the researcher. The classes are not limited by the training data. Any task and its classes can be verbalised.

2.3 Empirical analyses

2.3.1 Setup of empirical analyses: data and algorithms

To investigate the effects of transfer learning we analyse a diverse group of datasets, representing typical classification tasks which political scientists are interested in. The datasets vary in size, domain, unit of analysis, and task-specific research interest (see table 2). For all datasets, the overall task for human coders was to classify a text into one of multiple predefined classes of substantive political interest. Additional details on each dataset are provided in appendix A.

Different data pre-processing steps were tested. One objective during pre-processing is to align the classifier input more closely with the input human annotators receive. In some datasets, the unit of analysis for classification are individual quasi-sentences[7] extracted from longer speeches or party manifestos (Burst et al., 2020; Project, 2015). Human coders did, however, not interpret these quasi-sentences in isolation, but after reading the preceding (and following) text. Inspired by Bilbao-Jayo and Almeida (2018) we therefore test each algorithm with two types of inputs during hyperparameter search: only the single annotated quasi-sentence, or the quasi-sentence concatenated with its preceding and following sentence. See appendix E for other pre-processing steps for each algorithm.

Algorithms

Each dataset is analysed with the following algorithms:

- Classical algorithms: Support Vector Machines (SVM) and Logistic Regression – two widely used algorithms to represent classical approaches. For each classical algorithm we test two types of feature representations: TFIDF vectorization and average word embeddings (see appendix E4). Word embeddings provide a shallow form of 'language knowledge'.[8]

- A standard Transformer model: We use DeBERTaV3-base, which is an improved version of the original BERT trained on more

[7] A quasi-sentence is an entire sentence or a part of a sentence that represents one semantic unit. If one sentence contains two concepts of interest, it is split into two quasi-sentences.

[8] We use pre-trained GloVe embeddings (Pennington, Socher, & Manning, 2014) provided by the SpaCy library (en_core_web_lg-3.2.0 by Montani et al., 2022), a widely used type of word embedding (Rodriguez & Spirling, 2022).

Table 2.2: Key political datasets used in the analysis

Dataset	Task	Domain	Unit of Analysis	Includes Context?	Avg. Text length	Data Points Train / Test
Manifesto Corpus (Burst et al., 2020)	Classify text in 8 general topics	Party Manifestos	Quasi-sentences	Yes	116 characters (348 with context)	121570 all 88158 train 33412 test
Sentiment Economy News (Barberá et al., 2021)	Differentiate if economy is performing well or badly according to the text (2 classes)	News articles	News headline & first paragraphs	No	1624 cha.	3382 all 3000 train 382 test
US State of the Union Speeches (Project, 2015)	Classify text in policy topics (22 classes)	Presidential Speeches	Quasi-sentences	Yes	116 cha. (347 with context)	21641 all 15207 train 6434 test
US Supreme Court Cases (Project, 2014)	Classify text in policy topics (20 classes)	Law, summaries of court cases and rulings	Court case summaries (multiple paragraphs)	No	2456 cha.	7752 all 5236 train 2326 test
CoronaNet (Cheng et al., 2020)	Classify text in types of policy measures against COVID-19 (20 classes)	Texts from research assistants, news, governments	One or multiple sentences	No	297 cha.	48998 all 34298 train 14700 test
Manifesto stances towards the military (subset of Burst et al., 2020)	Identify stance towards the simple topic "military". (3 classes: positive/negative/unrelated).	Party Manifestos	Quasi-sentences	Yes	Similar to Manifesto Corpus above	13507 all 3970 train 9537 test
Manifesto stances towards protectionism (subset of Burst et al., 2020)	Identify stance towards the concept "protectionism" (3 classes: positive/negative/unrelated).	Party Manifestos	Quasi-sentences	Yes	Similar to Manifesto Corpus above	5878 all 2116 train 3762 test
Manifesto stances towards traditional morality (subset of Burst et al., 2020)	Identify stance towards the complex concept "traditional morality" (3 classes: positive/negative/unrelated).	Party Manifestos	Quasi-sentences	Yes	Similar to Manifesto Corpus above	7478 all 3188 train 4290 test

data, with a better pre-training objective than MLM and some architectural improvements (P. He et al., 2021, see appendix B2 for details).

- An NLI-Transformer: We fine-tune DeBERTaV3-base on 1.279.665 NLI hypothesis-context pairs from eight existing general-purpose NLI datasets ("BERT-NLI", see appendix B3).[9]

Converting political science tasks to NLI format and fine-tuning BERT-NLI

Specifically for fine-tuning BERT-NLI, the following steps were required. First, we read the codebook for each task and manually formulate one hypothesis corresponding to each class. For example, Barberá et al. (2021) asked coders to determine if a news article contains positive or negative indications on the performance of the U.S. economy. Based on the codebook, we therefore formulated the two class-hypotheses "The economy is performing well overall" and "The economy is performing badly overall".[10] Second, we optionally write a simple script to reformat the target texts to increase the natural language fit between the class-hypothesis and the target (con)text, if necessary.[11] Third, we fine-tune the general-purpose BERT-NLI model on e.g. 500 annotated texts from the Manifesto-military dataset. To this end, we match each text with the class-hypothesis we know to be 'true' based on the existing annotations and assign the label 'true'. In addition, we also match each text with one random 'not-true' class-hypothesis and assign the label 'neutral'. This avoids that BERT-NLI learns to only predict the class 'true' and provides a convenient means for data augmentation. The result is e.g. BERT-NLI-manifesto-military, which both 'knows' the general NLI task and the specific Manifesto-military task reformatted to NLI. Fourth, the fine-tuned model can then be applied to texts in a test set. As illustrated in figure 1, each test text is fed into BERT-NLI exactly N times, once with each of the N different class-hypotheses. The class for which the hypothesis is the most 'true' is selected.

Note that this approach allows us to further align the classifier input with the human annotator input: each human coder based their annotations on instructions in a codebook and with BERT-NLI we can

[9]The model is available at `https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c`

[10]In practice, we tested different hypothesis formulations during hyperparameter search, see appendix B and E.

[11]For some tasks, we found that reformatting the context to 'The quote: "context" ' and formulating the hypotheses as 'The quote is about …' increases the natural language fit between hypothesis and context, which increases performance (see appendix B). The literature uses less natural formulations like 'It is about …' (Yin et al., 2019).

provide these coding instructions to the model via the class-hypotheses (see 'label verbalisation' above and appendix B).

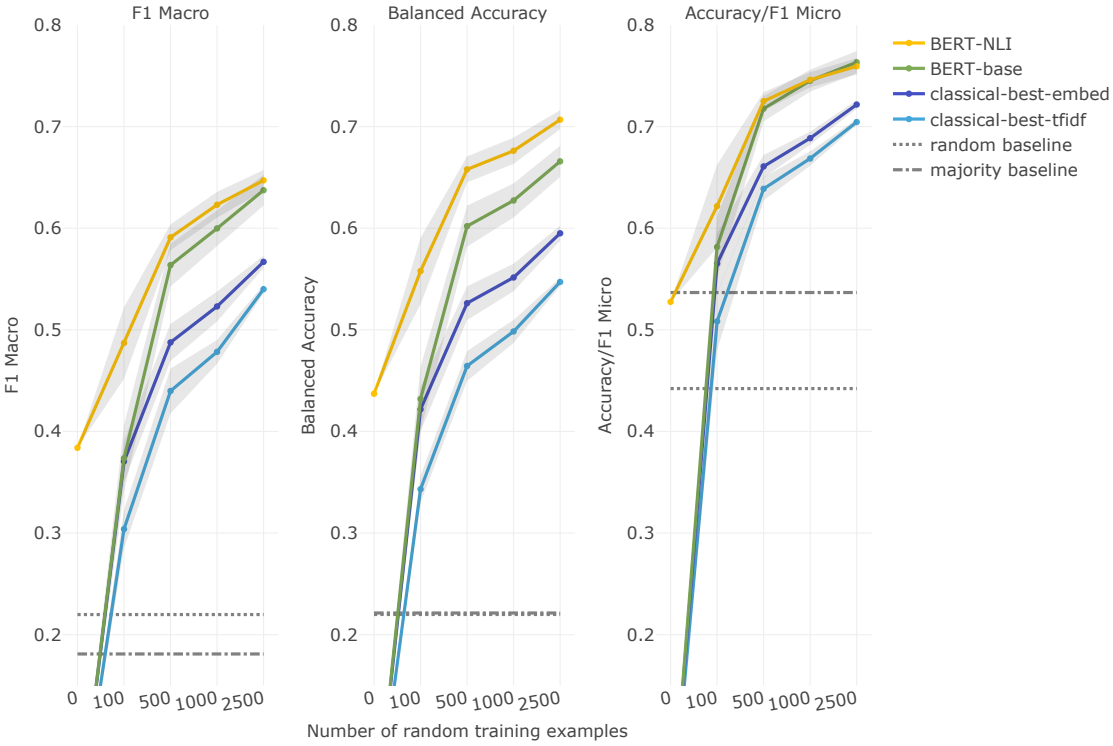
Comparative analysis pipeline and metrics

The objective of our analysis is to determine how much data, and therefore annotation labour, is necessary to obtain a desired level of performance on diverse classification tasks and imbalanced data. To ensure comparability and reproducibility across datasets and algorithms, each dataset is analysed based on the same script: the random training sample size is successively increased from 0 to 10 000 texts, hyperparameters are tuned on a validation set, final performance is tested on a holdout test set. We assess uncertainty by taking three random training samples and report standard deviation (see appendix C).

We evaluate each model and task with multiple metrics (following the implementations by Pedregosa et al., 2011). Firstly, accuracy counts the overall fraction of correct predictions (and is equivalent to F1 Micro). The disadvantage of accuracy is that it overestimates the performance of classifiers overpredicting majority classes and neglecting minority classes. On three of our tasks, a baseline model that only predicts the majority class would already achieve above 90% accuracy due to high data imbalance. We assume that in most social science use-cases, all classes included in a task are of roughly similar importance, making accuracy a misleading metric for performance. Secondly, balanced accuracy calculates accuracy for each class separately and then takes the average of each per-class accuracy score (equivalent to 'Recall Macro'). This gives equal weight to all classes independently of their size and is a more suitable metric, assuming that classes have similar substantive value. A characteristic of balanced accuracy is that it is higher for classifiers with less false negatives (high 'Recall') but does not properly account for false positives (risk of lower 'Precision'). Balanced accuracy empirically favours classifiers that predict many minority classes well but perform less well on a few majority classes (appendix D1). Thirdly, F1 Macro is a metric that tries to remedy this issue. It is the harmonic mean of Precision and Recall and gives equal weight to all classes independently of their size. Appendix D provides a more detailed empirical discussion and data, including other metrics like Cohen's Kappa. We conclude that F1 Macro is the most adequate metric for many social science use-cases of supervised machine learning and we therefore use it as the primary metric in this paper, while also reporting other metrics.[12]

[12]Note that the importance of different classes might vary in different substantive

Figure 2.2: Average performance across eight tasks vs. training data size



The 'classical-best' lines display the results from either the SVM or Logistic Regression, whichever is better. Note that four datasets contain more than 2500 data points, see figure 3.

2.3.2 Empirical results

Figure 2 displays the aggregate average scores across all datasets. Figure 3 displays the results per dataset (see appendix D for detailed metrics). We focus on two main aspects across tasks: overall data efficiency and ability to handle imbalanced data.

Regarding data efficiency, deep transfer learning models perform significantly better with less data than classical models across all tasks. The results show that BERT-NLI outperforms the classical models with

research projects and researchers can make more nuanced decisions on the weight they attribute to different classes.

TF-IDF by 10.7 to 18.3 percentage points on average (F1 Macro) when 100 to 2500 annotated data points are available (7.9 to 12.4 with BERT-base). Classical models can be improved by leveraging shallow 'language knowledge' from averaged word embeddings, but a performance difference of 8.0 to 11.7 F1 Macro remains (0.4 to 7.7 with BERT-base). The results indicate that BERT-NLI achieves similar average F1 Macro performance with 500 data points as the classical models with around 5000 data points.[13] The performance difference remains, as larger amounts of data are sampled (5000 – 10 000, see figure 3 and appendix D3) and applies across domains, units of analysis and tasks.

Moreover, the more transfer learning components a model is using, the better it becomes at handling imbalanced data. We demonstrate this by comparing accuracy/F1 Micro to F1 Macro averaged across the data intervals 100 to 2500. Higher improvements with F1 Macro indicate an improved ability to handle imbalanced data. When 'shallow language knowledge' with word embeddings is added to classical model instead of TFIDF, F1 Macro is increased by +4.6 percentage points, while accuracy/F1 Micro is only increased by +2.9 – a +1.7 higher improvement for F1 Macro. With BERT-base and its 'deep language knowledge', the improvement over classical TFIDF is +7.2 with accuracy/F1 Micro and +10.3 with F1 Macro – a +3.1 higher improvement for F1 Macro. With BERT-NLI and its additional 'task knowledge', the improvement is +8.3 with accuracy/F1 Micro and 14.6 with F1 Macro – a +6.3 higher improvement for F1 Macro. The higher F1 Macro score improvements compared to accuracy/F1 Micro indicates that transfer learning reduces reliance on majority classes. Good classifiers should perform similarly across all classes a researcher is interested in. Appendix D1 provides additional data demonstrating that, when more transfer learning components are added, the performance on different classes becomes less varied.

This has two main reasons: First, both BERT variants (and word embeddings) require fewer examples for the words used in minority classes thanks to their prior representations of e.g. synonyms and semantic similarities of texts ('language knowledge'). Second, BERT-NLI performs better on F1 Macro and especially balanced accuracy and its performance across classes is least varied. Its prior 'task knowledge'

[13]Note that the results above 2500 data points are harder to compare, as only 4 datasets have enough data for the data intervals of 5000 or more. This statement is therefore based on the performance for 4 datasets (see appendix D) as well as the overall trendline for all 8 datasets.

further reduces the need for data for smaller classes. In appendix D1 we show empirically that the comparatively high performance of BERT-NLI on balanced accuracy is due to higher performance on many smaller classes compared to few majority classes. BERT-NLI can already predict a class without a single class example in the data ('zero-shot classification'). It does not need to learn each class for the new task since it uses the universal NLI task where classes are expressed in hypotheses verbalising the codebook. This capability is also illustrated in figures 2 and 3 by the metrics with zero training examples.

Note that our metrics are based on fully random training data samples, which do not always contain examples for all classes, especially for datasets with many classes. This simulates a typical challenge social scientists are facing, where random sampling is common and even advanced sampling techniques like active learning require an initial random sampling step (Miller et al., 2020). Transfer learning and especially prior 'task knowledge' can therefore become another tool in our toolbox to address the issue of imbalanced data. Also note that the values for accuracy/F1 Micro are significantly higher than for F1 Macro for all models and only reporting accuracy/F1 Micro provides a misleading picture of actual performance on imbalanced data.

How to choose between BERT-base and BERT-NLI? The main criteria are the amount of training data and the degree of data imbalance. BERT-NLI is useful in situations where little and very imbalanced data is available $<\approx 1000$. As more data becomes available to learn the new task (and minority classes) from scratch, it seems advisable to use the simpler BERT-base model given the converging performance $>\approx 2000$. BERT-NLI has a tendency to perform better on (many) minority classes, while performing less well on (few) majority classes – which can be good or bad, depending on the use-case (see appendix D). Another dataset characteristic that can influence the value of BERT-NLI is concept complexity. BERT-NLI seems to work better when concepts are measured that can be clearly expressed in the hypotheses. For example, it performs particularly well on the Manifesto-military task, measuring the stance towards the comparatively simple topic 'military'. At the same time, it performs comparatively less well on Manifesto-morality where the complex concept 'traditional morality' is measured, which covers diverse sub-dimensions from traditional family values, religious moral values to unclear concepts like 'unseemly behaviour'. We assume that it is harder for BERT-NLI to map the simple language in the hypothesis

to complex concepts. We discuss other factors that can influence the performance of BERT-NLI in appendix B4.

Lastly, we observe that hyperparameters and text pre-processing can have an important impact on performance for all models. For example, while BERT-base models are normally trained for less than 10 epochs, we find that training for up to 100 epochs increases performance on small datasets (see appendix E3 for a systematic study on hyperparameters). Moreover, regarding pre-processing, if the unit of analysis are quasi-sentences, including the preceding and following sentence during pre-processing systematically increases performance for all models (appendix E1); the value of word embeddings can be increased by reweighting the averaged embeddings and selecting more important words with part-of-speech tagging (appendix E4); and the performance of BERT-NLI can be improved through simple pre-processing steps (appendix B5).

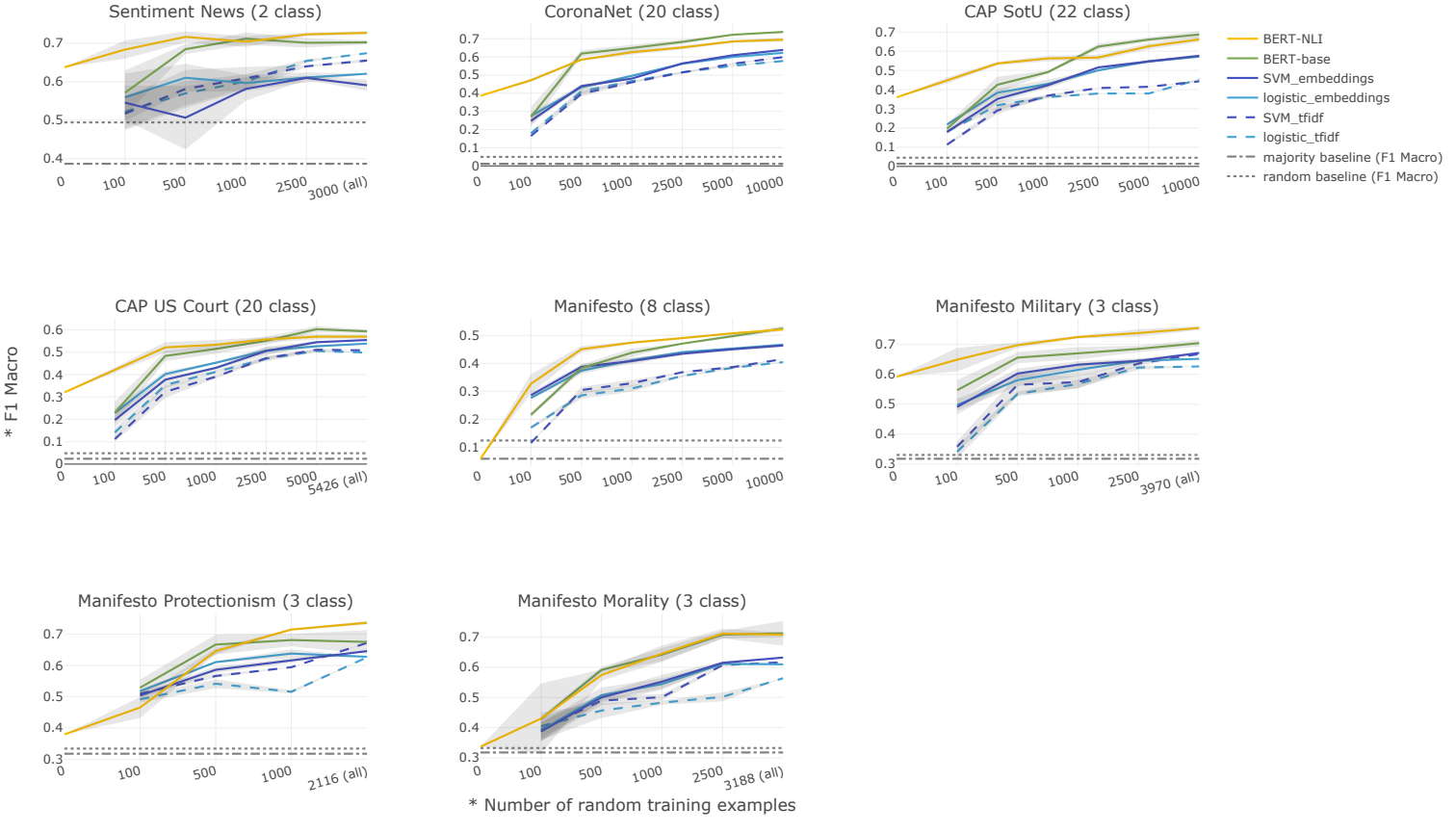
2.4 Discussion of limitations

While deep transfer learning leads to high classification performance, several limitations need to be discussed. First, deep learning models are computationally slow and require specific hardware. BERT-like Transformers take several minutes to several hours to fine-tune on a high-performance GPU, while a classical model can be trained in minutes on a laptop CPU. To help alleviate this limitation, we share our experience for accessing GPUs (appendix F) and choosing the right hyperparameters (appendix E3). Our extensive hyperparameter experiments indicate that a set of standard hyperparameters performs well across tasks and data sizes and researchers can refer to these default values to reduce computational costs.

Moreover, using BERT requires learning new software libraries. Luckily, there are relatively easy to use open-source libraries like Hugging Face Transformers, which only require a moderate understanding of Python and no more than secondary education in math (Wolf et al., 2020).[14] Furthermore, specifically for BERT-NLI, we share our models and code. We provide several BERT-NLI models used in this paper with state-of-the-art performance on established NLI benchmarks. We invite researchers to copy and adapt our models and code to their own

[14] Hugging Face also provides a beginner-friendly course: `https://huggingface.co/course/chapter1/1`

Figure 2.3: Performance per task vs. training data size (F1 Macro)



*Note that the x- and y-axis scales can differ across datasets

datasets.[15]

An additional disadvantage specifically of NLI is its reliance on human annotated NLI data, which is abundantly available in English, but less so in other languages. We also provide a multilingual BERT-NLI model pre-trained on 100 languages, but we expect it to perform less well than English-only models (appendix B).[16] There are several other techniques for leveraging 'prior task knowledge' which do not rely on human annotated data and could be explored in future research (Brown et al., 2020; Schick & Schütze, 2021a).

Lastly, model (pre-)training can introduce biases and impact the validity of outputs. There is a broad literature on bias in deep learning models (Blodgett, Barocas, Daumé III, & Wallach, 2020) and this most likely extends to political bias and NLI. It is possible, for example, that the hypotheses "The US is trustworthy" and "China is trustworthy" will result in different outputs for semantically equal inputs as one actor might have been mentioned more often in a negative context than others during (pre-)training. Political bias in deep learning is an important subject for future research. Moreover, the 'black box' nature of deep learning models makes them harder to interpret. This becomes problematic when researchers want to understand why exactly a model has made a certain classification. There are some open-source libraries such as Captum[17] which can partly alleviate this issue by extracting the importance of specific features (words) for a classification decision to enable interpretations. More generally, whether the supervised machine learning pipeline used for a specific new research question is internally and externally valid is an important additional assessment for substantive research projects (Baden et al., 2022).

2.5 Conclusion and outlook

Lack of training data is a major hurdle for researchers who consider using supervised machine learning. This paper outlined how deep transfer learning can lower this barrier. Transformers like BERT can store information on statistical language patterns ('language knowledge') and

[15] NLI models are available at `https://huggingface.co/MoritzLaurer`; an easy-to-use Jupyter notebook to train your own BERT-NLI model is available at: `https://github.com/MoritzLaurer/less-annotating-with-bert-nli`

[16] `https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli`

[17] `https://github.com/pytorch/captum`

they can be trained on a universal task like NLI to help them learn downstream tasks and classes more quickly ('task knowledge'). In contrast, classical models need to learn language and tasks from scratch with the training data as the only source of information for any new task.

We systematically test the effect of transfer learning on a range of eight tasks from five widely used political science datasets with varying size, domain, unit of analysis, and task-specific research interest. Across these eight tasks, BERT-NLI trained on 100 to 2500 data points performs on average 10.7 to 18.3 percentage points better than classical models with TF-IDF vectorization (F1 Macro). We also show that leveraging the shallow 'language knowledge' of averaged word embeddings with classical models improves performance compared to TF-IDF, but the difference to BERT-NLI is still large (8.0 to 11.7 F1 Macro). Our study indicates that BERT-NLI trained on 500 data points achieves similar average F1 Macro performance as classical models with around 5000 data points. Moreover, transfer learning works particularly well for imbalanced data, as it reduces the data requirements for minority classes. We also provide advice on when to use BERT-NLI and when using a simpler BERT-base model is advisable. Researchers can use our results as a rough indicator for how much annotation labour their task could require with different methods.

Based on these empirical findings, we believe that deep transfer learning has great potential for making supervised machine learning a more valuable tool for social science research. As most research projects tackle new research questions which require new data for different tasks on mostly imbalanced data, the reduction of data requirements is a substantial benefit. Moreover, this enables researchers to spend more time on ensuring data quality rather than quantity and carefully creating test data for ensuring the validity of models. Accurate models combined with high quality datasets directly contribute to the validity of computational methods.

There are many important directions for future research this paper could not cover. This paper used random sampling for obtaining training data. Active learning can further reduce the number of required annotated examples (Miller et al., 2020). In fact, combinations of active learning and BERT-NLI are promising, as the zero-shot classification capabilities of BERT-NLI can be used in the first sampling round. Moreover, issues of political bias and validity need to be investigated

further. Computational social scientists should become a more active part of the debate on (political) bias and validity in the machine learning community.

Lastly, we believe that transfer learning has great potential for enabling the sharing and reusing of data and models in the computational social sciences. Datasets are traditionally mostly designed for one specific research question and fine-tuned models can hardly be reused in other research projects. Transfer learning in general and universal tasks in particular can help break these silos. Computational social scientists with a 'transfer learning mindset' could create general purpose datasets and models designed for a wider variety of use cases. Transfer learning opens many new venues for sharing and reuse which have yet to be explored.

Appendix

The extensive appendix is available online via the published version of the paper: *Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023). 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*

Chapter 3

Lowering the Language Barrier

Investigating Deep Transfer Learning and Machine Translation for Multilingual Analyses of Political Text

Abstract. *The social science toolkit for computational text analysis is still very much in the making. We know surprisingly little about how to produce valid insights from large amounts of multilingual texts for comparative social science research. In this paper, we test several recent innovations from deep transfer learning to help advance the computational toolkit for social science research in multilingual settings. We investigate the extent to which prior language and task knowledge stored in the parameters of modern language models is useful for enabling multilingual research; we investigate the extent to which these algorithms can be fruitfully combined with machine translation; and we investigate whether these methods are accurate, practical and valid in multilingual settings – three essential conditions for lowering the language barrier in practice. We use two datasets with texts in 12 languages from 27 countries for our investigation. Our analysis shows, that, based on these innovations, supervised machine learning can produce substantively meaningful outputs. Our BERT-NLI model trained on only 674 or 1,674 texts in only one or two languages can validly predict political party families' stances towards immigration in eight other languages and ten other countries.*

Paper published as: *Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023). 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*

3.1 Introduction

While computational text analysis methods have exploded in popularity over the past decade, our computational toolkit is still very much in the making. We know surprisingly little about how to produce valid measurements with computational methods that are directly useful for substantive comparative social science research. This issue is even worse in multilingual settings (Baden et al., 2022). One important reason for this is the language barrier: People do not speak many languages besides their mother tongue and if they do, their language knowledge is limited to a few dominant languages like English (Eurobarometer, 2012). This is also reflected in the computational text analysis literature, where most research focuses on English, while multilingual tools are lacking (Baden et al., 2022).

For supervised machine learning, this language barrier leads to several important challenges which limit its utility for comparative multilingual research. Supervised text classification traditionally requires large amounts of training data for classifiers to obtain a useful level of accuracy. This issue is aggravated in multilingual research settings, where even more data and language knowledge are required for the different languages, countries and cultures that are relevant for comparative research questions. Researchers regularly face the question whether they have the human resources and computational tools available for meaningful comparative research on large amounts of multilingual texts. The answer to this question is probably too often: No.

This paper investigates different computational methods for lowering this language barrier. We investigate the extent to which deep transfer learning can help address the issue, by leveraging 'prior language and task knowledge' stored in the parameters of modern language models (Ruder, 2019); we investigate the extent to which these models can be combined with machine translation to further lower this barrier; and we investigate whether these methods are not only accurate but also practical and valid – three important criteria for lowering the language barrier in practice.

We show empirically that transfer learning models are useful tools for comparative multilingual research. We design two analyses to systematically compare different methods on two datasets with texts in 12 languages from 27 countries: a simple topic classification task based on the Manifesto corpus (Burst et al., 2020) and a complex task on stances towards immigration (Zobel & Lehmann, 2018). We empirically

compare the performance and usability of many different combinations of methods on both datasets. Lastly, we show that certain transfer learning models do not only produce accurate, but also valid outputs using one prominent use-case. Our best model (BERT-NLI) fine-tuned on only 674 or 1,674 texts in only one or two languages can validly predict political party families' stances towards immigration in eight other languages and ten other countries. We open-source our models, a custom dataset of 2.7 million NLI texts in 26 languages spoken by around 4 billion people, as well as the full reproduction code.[1] We conclude by discussing limitations of different methods and opportunities for future research. Through our investigation we hope to contribute to the 'accumulation of methodological knowledge and guidance, as well as the strategic improvement of available tools' (Baden et al., 2022, p. 14). Based on the methodological findings in this paper, we hope to provide guidance for researchers to help them tackle a key challenge of social science research: creating high quality measurements from multilingual text for answering complex substantive questions.

3.2 Existing literature

3.2.1 Machine translation and vector representation approaches

The challenge of computationally analysing multilingual texts has been studied with different methods and objectives in both the social sciences and computer linguistics. In the social sciences (for a good overview, see Licht & Lind, 2023), the most widely used approach is *machine translation (MT)*. For this approach, multilingual texts are machine translated into a single anchor language, which classical algorithms can process, and researchers can understand. Several authors have demonstrated the utility of MT for computational text analyses: Lucas et al. (2015) use MT to estimate topic models related to Jihadi Fatwas or reactions to the Snowden revelations in China; de Vries et al. (2018) show that topic modelling leads to comparable results on machine translated and human translated texts; Windsor, Cupit, and Windsor (2019) argue that applying a sentiment dictionary to UN speeches leads to similar results on machine and human translated texts; Düpont and Rachuj

[1] Our models and dataset are available at `https://huggingface.co/MoritzLaurer`, our reproduction code is available at `https://github.com/MoritzLaurer/language-barrier-multilingual-transfer`

(2022) use machine translated party manifestos from 19 countries to study policy diffusion across countries with textual similarity measures. Regarding supervised machine learning and text classification (the focus of this paper) Courtney, Breen, McMenamin, and McNulty (2020) show that manual annotations on machine translated texts are comparable to annotations on original texts in an annotator's mother tongue, and Lind, Heidenreich, Kralj, and Boomgaarden (2021) use machine translation combined with a dictionary to enable supervised classification.

As an alternative (or supplement) to MT, the computer linguistics literature has introduced *multilingual vector representation approaches* (Conneau et al., 2020; Ruder, Vulić, & Søgaard, 2019). While MT approaches align text by translating all texts into the same language, multilingual representation approaches align texts by representing all texts in the same vector space. These vector-based approaches try to take texts from different languages and convert semantically similar texts to similar vectors independently of the original language. Many different approaches using multilingual vector representations exist. The first popular approach was multilingual word embeddings. It extends the idea of classical word embeddings (Mikolov et al., 2013), only that it also represents tokens from different languages in the same vector space (Ruder et al., 2019). For example, the tokens 'love' and 'amour' would be represented by similar vectors, while 'Auto' would be represented by a relatively distant vector.

A few years later, the deep learning literature further improved on (multilingual) embedding approaches. Researchers introduced monolingual transformers like BERT (Devlin et al., 2019), which were soon extended to become multilingual (Conneau et al., 2020; Conneau & Lample, 2019). A monolingual BERT model consists of a vocabulary layer of around 30,000 (English) word vectors followed by multiple layers of additional vectors designed to create contextualised vector representations. Concretely, when a text is provided to BERT, the text is (a) tokenized, (b) each token is converted to a word vector representation stored in the vocabulary layer, (c) the following layers change the representation of each word vector depending on its surrounding words and (d) the last task-specific layer produces an output, e.g. probabilities for different classes. The vectors in the different layers are tuned during a pre-training phase with a self-supervised task like Masked Language Modelling (MLM) (Devlin et al., 2019). To make these models multilingual, the only main changes are the pre-training data and the initial

vocabulary layer: First, instead of self-supervised pre-training on English texts, they are pretrained on more texts from up to 100 languages at the same time; second, the vocabulary layer is extended from around 30,000 tokens to around 250,000 tokens to account for the wider vocabulary necessary for many languages and scripts. The multilingual BERT (mBERT) then works essentially the same as a monolingual BERT and can be fine-tuned on any of the 100 languages it has been pre-trained on (Conneau et al., 2020; P. He et al., 2021).

What makes multilingual representation models an interesting alternative to MT is their capability of 'cross-lingual transfer learning' (Conneau et al., 2020). mBERT can be fine-tuned to perform a task with data from one (or more) source language(s) and then perform the same task on any other target language it has learned during pre-training - without requiring fine-tuning data from the target languages.

Many different variants of BERT exist and can be used to classify multilingual texts. First, a normal pre-trained mBERT can directly be fine-tuned on a classification task. Second, a multilingual Sentence-BERT can be used to generate multilingual sentence embeddings (Reimers & Gurevych, 2020). These sentence embeddings can then be used as the training input for a classical algorithm like a Logistic Regression – a method that has previously been used by political scientists (Licht, 2023). This approach is similar to feeding averaged word embeddings into a Logistic Regression, only that sentence embeddings are better at approximating the meaning of texts longer than individual words (Reimers & Gurevych, 2019). The main advantage of this approach for classification is that Sentence-BERT is only used to produce embeddings without being fine-tuned and then only a simple algorithm like Logistic Regression is fine-tuned on these embeddings. This strongly reduces computation costs. Third, multilingual universal classifiers like mBERT-NLI can be used for classification. With BERT-NLI algorithms, the researcher first formulates a 'class-hypothesis' for each label that verbalises the underlying classification task. 'The text is positive' and 'The text is negative' could be two hypotheses for a binary sentiment classification task. Each target text in the corpus is then fed into mBERT-NLI together with the class-hypotheses and the model tries to predict which class-hypothesis is 'truest' for the respective text. This approach has been shown to perform very well in monolingual use-cases, when only little and imbalanced data is available (Laurer, Van Atteveldt, Casas, & Welbers, 2023a; S. Wang et al., 2021).

3.2.2 Gaps and research questions

Given the relative novelty of multilingual transformers like mBERT, we do not know enough about the advantages and disadvantages of different approaches for multilingual social science research.

First, most variations of mBERT have not yet been studied for social science tasks. Only the value of multilingual Sentence-BERT for classifying texts from party manifestos has been investigated in one paper (Licht, 2023). The author uses a pre-trained multilingual Sentence-BERT to create multilingual sentence embeddings as the input for training a classical regression model. They do not fine-tune the BERT model. The recent political science literature shows that fine-tuning BERT outperforms classical algorithms for text classification on monolingual data (Bestvater & Monroe, 2022; Terechshenko et al., 2020; Widmann & Wich, 2022). In a multilingual setting however, it is unclear how fine-tuning an mBERT model compares to other approaches. The same is the case for universal classification approaches like BERT-NLI, which have only been applied to English data (Laurer et al., 2023a). This paper investigates: What are the advantages and disadvantages of different algorithms in multilingual settings in terms of performance and usability for social science tasks?

Secondly, multilingual vector representation approaches and MT can be treated as alternatives, but also as supplements. The main benefit of mBERT is that they can ingest texts in up to 100 languages at the same time. A supplementary benefit of MT is that one text can be translated into many different languages. This also means that one text in one language can be machine translated to multiple texts in multiple languages and mBERT can use all texts for training. Barriere and Balahur (2020), for example, take tweets in five languages annotated for sentiment and translate each tweet to the respective other four languages. They then use the combined original and translated tweets to train an mBERT model and achieve higher performance than monolingual BERT trained only on the monolingual data. Bornea, Pan, Rosenthal, Florian, and Sil (2021) use a similar approach for question answering tasks. The assumption is that translating one text into multiple languages increases quantity and variety of data, which can increase performance. MT could therefore provide a convenient means for data augmentation (Li, Hou, & Che, 2022) when combined with mBERT and when limited amounts of data are available. At the same time, this combined approach risks creating noisy data. This paper investigates: What are the advantages

and disadvantages of combining MT with mBERT for social science tasks?

Third, our knowledge of the specific challenges for validating multilingual text classifiers is limited. Supervised machine learning has some in-built validation via the held-out test set, but risks of low external validity and biases remain. This is even more severe in multilingual settings. Aligning text inputs from different languages is only one challenge, for which the two main solutions were outlined above. Accounting for cultural differences that cannot simply be input-aligned is a second important challenge amplified in multilingual settings. Lexically identical texts can have a dramatically different meaning depending on the language and country of origin of the speaker. Take the phrase 'We are proud of the achievements of the political right in our country'. If uttered by a German politician, this phrase would evoke a specific meaning for German readers. A German reader would probably place the politician at the far right of the political spectrum, as the identification as "politically right" ("politisch rechts") is not accepted in the German mainstream, given the term's association with Nazi Germany. At the same time, if the exact same phrase is uttered in French by a French politician ("la droite"), French readers could reasonably place the politician only slightly right to the political center, given the different national history of the term. Our toolkit for addressing this second challenge is even less developed and it cannot be analysed with standard aggregate metrics. This paper investigates: How can the validity in multilingual text classification be assessed and how do different approaches impact validity?

Answering these questions can contribute to the broader empirical challenge of lowering the language barrier for comparative research. Only if supervised machine learning is accurate, usable, and valid in multilingual settings can it be a useful tool for substantive comparative social science research.

3.3 Methodology

To address the research questions outlined above, we apply combinations of different multilingual text analysis approaches to several practical multilingual research scenarios. The first analysis focuses on performance and usability while the second analysis focuses on validity.

3.3.1 Analysis 1: Comparing machine translation and (m)BERT

Scenarios. We analyse two practical scenarios where the language barrier hinders comparative research. First, in the *low-resource scenario* we assume that a research team has annotated texts for a task in one source language and wants to apply the task to texts in another target language for which they do not have annotated data. We choose English as the source language, as it is a widely spoken and data-rich language ("one language" scenario in figure 3.1). Second, in the *higher-resource scenario* a research team has annotated texts for a task in multiple languages and they want to build a classifier that is robust across all these languages ("many languages" scenario in figure 3.1). For both scenarios, we assume that the research team has limited resources and therefore assume that 500 annotated texts are available per source language: In total 500 texts in the first scenario and 500 per language in the second scenario. To each of these two scenarios, we apply three different machine translation approaches (second column in figure 3.1) and four different algorithms (third column in figure 3.1).

Datasets. We selected two datasets that contain multilingual data and tasks that are highly relevant for comparative political science research. Firstly, we choose the Manifesto Corpus and its topic identification task (Burst et al., 2020). The dataset is widely used in the comparative politics literature and it is one of the few datasets with a harmonised category scheme across multiple languages including non-Western languages. For our analysis, we choose texts in the following languages: English, French, German, Korean, Russian, Spanish, and Turkish. Selection criteria for these languages were the global number of native speakers, sufficient data in the corpus, and diversity in terms of culture and scripts. As the main task, we chose the classification of quasi-sentences[2] into general topical domains. More specifically, a model needs to learn to classify a text into one of the following eight topical categories: economy, external relations, fabric of society, freedom and democracy, political system, social groups, welfare and quality of life, other. The Manifesto Corpus also contains more fine-grained categories, but many of them contain only very little data for non-Western

[2]A quasi-sentence is an entire sentence or a part of a sentence that represents one semantic unit. If one sentence contains two concepts of interest, it is split into two quasi-sentences (Merz, Regel, & Lewandowski, 2016).

languages which are often neglected in the text-as-data literature.[3]

Secondly, we choose the PImPo dataset on 'Parties' Immigration and Integration Positions' (Zobel & Lehmann, 2018), which represents a complex stance detection task. PImPo analyses political parties' stances towards two concepts: immigration and integration. For the purpose of our analysis, we only analyse stances towards immigration. The task for the model is to identify if a given text is "supportive of immigration", "sceptical towards immigration", "neutral towards immigration", or "not about immigration" (four-class classification). All texts that are about "integration" are added to the "not about immigration" category, creating hard edge cases for the model to distinguish from texts about immigration. The dataset was created in a large-scale crowd-coding exercise, where crowd-workers annotated over 200,000 quasi-sentences from party manifestos. The data is available in nine languages from 14 (unfortunately only Western) countries[4]: Danish, Dutch, English, Finnish, French, German, Norwegian, Spanish, and Swedish.[5]

By choosing these two datasets, we test our approaches on both a relatively simple topic identification task with eight classes and a more complex stance detection task that requires identifying a specific topic (immigration vs. integration or other unrelated texts) and a stance towards this topic (positive/negative/neutral). We limit our analysis to these two datasets due to the lack of publicly available, high quality and high quantity datasets with multiple (non-Western) languages that would have fit our comparative high/low-resource comparative pipeline. Additional details on both datasets are available in appendix A.[6]

Algorithms. We systematically compare several different algorithms:

[3]We had originally planned on including other non-Western languages like Japanese, but eventually did not include them due to lack of data for minority classes and to limit the complexity of the study.

[4]The countries are Sweden, Norway, Denmark, Finland, Netherlands, Spain, Germany, Austria, Switzerland, Ireland, United States, Canada, Australia, New Zealand.

[5]The PImPo authors mention ten languages and probably refer to a few Gaelic texts in the Irish data as the tenth language. As we could only identify 64 Gaelic texts and none of them were about the topics of interest, we excluded Gaelic from our analysis.

[6]The PImPo dataset, for example, is highly imbalanced (96% of texts belong to the 'no topic' class). We therefore downsample the 'no topic' class to maximum 2,000 texts per language. Methods for addressing data imbalance have been addressed in other research (Miller et al., 2020) and we leave more advanced sampling techniques to future research.

- A Logistic Regression with Bag-of-Words input and TF-IDF vectorization, representing classical computational text analysis approaches as a baseline (Osnabrügge et al., 2021).
- A Sentence-BERT model which creates sentence embeddings for each text. These embeddings are then used as the training input for a Logistic Regression instead of TF-IDF (Licht, 2023).
- A standard BERT-base model directly trained on the (multilingual) texts (Devlin et al., 2019; P. He et al., 2021).
- A BERT-NLI model trained following the approach from (Laurer et al., 2023a; S. Wang et al., 2021). For training our multilingual BERT-NLI model, we created a custom dataset of 2.7 million NLI text pairs machine translated to 26 languages spoken by roughly 4 billion people. With this dataset, we address the challenge that NLI data is mostly available in English.[7]

We choose each of these algorithms for the following reasons: Logistic Regression with TF-IDF input is a baseline method, that is widely used in the social sciences; the Sentence-BERT model was used in one of the first prominent social science papers that uses transformers (Licht, 2023); standard BERT-base is the main baseline for papers investigating transfer learning; and BERT-NLI is a newer method that tries to push the limits of transfer learning, also proposed in a recent social science paper (Laurer et al., 2023a). For each of the three BERT variants, we test both a monolingual English variant and a multilingual variant. The exact BERT variants (DeBERTaV3 and MPNet) are detailed in appendix C. Additional details on preprocessing, hyperparameter tuning and measures to handle randomness are provided in appendix D3.

Machine translation and data augmentation. As discussed above, MT can be used as an alternative or supplement to multilingual representation approaches. First, we test (multilingual) algorithms with the original, non-translated texts ("no-MT"). Second, we test each (monolingual) algorithm with texts that we machine-translated to the anchor language (English, 'translate2anchor'). Third, we translate each text to all other languages ('translate2many') and train the algorithms on this augmented input. Each of these three methods are tested for both the low-resource/one-language scenario and the high-resource/many languages scenario (see figure 3.1). All translations for this paper were implemented with open-source transformer-based machine translation

[7]The dataset is available for download and with more information at: `https://huggingface.co/datasets/MoritzLaurer/multilingual-NLI-26lang-2mil7`

models by Fan et al. (2020). Additional details are provided in appendix B. We control against potential negative effects of noisy data from these augmentation approaches by testing on human annotated (non-augmented) test-sets.

This combination of scenarios, algorithms and MT strategies enables a wide variety of combined approaches: 13 for the low-resource scenario and 21 for the higher-resource scenario. Appendix D1 displays all possible approaches and figure 3.1 provides a simplified overview. The overall objective of this analysis is to provide a systematic review of which combination of methods performs best with limited amounts of data, which social scientists typically have at their disposal. For this first analysis, we use the following criteria for evaluating each approach: (1) Aggregate performance on the standard metrics accuracy and F1-macro;[8] (2) how (un)equally an algorithm performs on different languages, measured by the standard deviation of performance across all languages; (3) ease-of-use for implementing the approach.

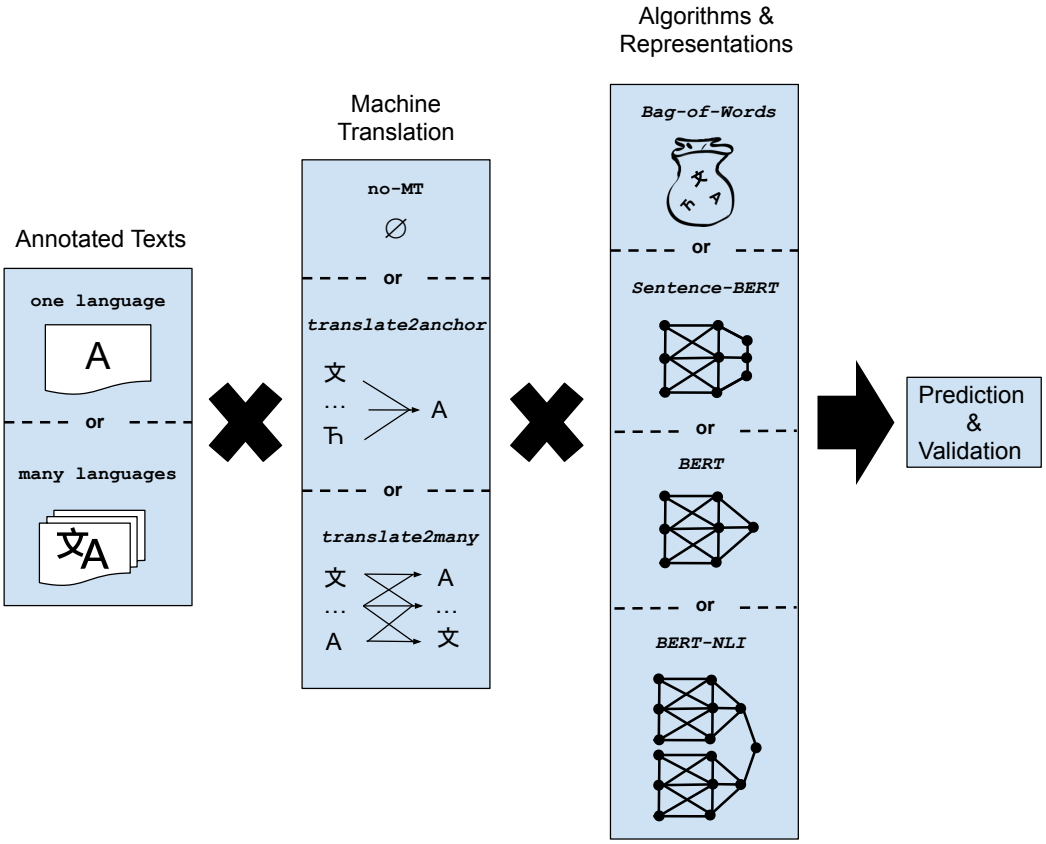
3.3.2 Analysis 2: Testing the validity of different computational approaches

The first analysis focuses on standard performance metrics. While these metrics provide good indicators of classification performance, the actual main interest for social scientists is the utility of an approach for creating substantive insights about social reality. This second analysis therefore focuses on validity: To what extent do (multilingual) supervised classifiers actually measure what they are intended to measure?

Validation and the PImPo dataset. There are many different definitions and sub-categories of validity and the use of the term varies across and within disciplines (Newton, 2012; Newton & Baird, 2016). For the purpose of this paper, we assess two types of validity based on the practical definitions used by Zobel and Lehmann (2018). First, we assess face-validity, asking: is an algorithm's prediction about a social phenomenon in line with theoretically founded expectations about the social phenomenon? Second, we assess convergent validity, asking: is an

[8]We base our choice of metrics on the empirical comparison of metrics in (Laurer et al., 2023a). Accuracy is an easily interpretable metric, but exaggerates the performance of classifiers overpredicting majority classes. F1-macro is the harmonic mean of precision and recall and attributes equal value to all classes, independently of their size. Assuming that each class has the same substantive value independently of their size, F1-macro is a better reflection of classifier performance on imbalanced datasets.

Figure 3.1: Tested approaches with possible combinations of different methods



algorithm's prediction about a social phenomenon correlated with a different external measurement of the same social phenomenon? The social phenomenon we analyse for validation is political parties' positions on immigration. As an existing external measurement of this phenomenon, we use the PImPo dataset from the first analysis. We choose the PImPo dataset for validation, since Lehmann and Zobel themselves provide validity tests for their crowd-coded data, which we can build upon (2018, p. 1070). Through these additional validity tests, we go beyond standard machine learning metrics and conventional definitions of 'performance'.

Selected approaches for the second analysis. As we will show below, the first analysis indicates that BERT variants clearly outperform

the classical Logistic Regression with TF-IDF and that MT for data augmentation only partly improves performance, while being relatively unpractical. Based on these observations and following the principle of parsimony, we exclude the classical Logistic Regression with TF-IDF and MT for data augmentation from this second analysis. Instead, we also tested two sizes of BERT in this second analysis: based-sized, which is the standard size from the first analysis and can be easily trained on publicly available GPUs; and large-sized, which can still be trained on publicly available GPUs, but more slowly. We include larger BERT models in this analysis to tests how advances in algorithms and hardware will impact validity in the coming years. New transformers are published every year and this increase in size is a rough proxy of possible improvements with future smaller language model variants.

Similar to the first analysis we also analyse each approach with an increasing number of languages for training, in order to study the impact of increasing multilingual resources: (1) only English training data (low-resource, monolingual scenario); (2) English and German (bilingual scenario); (3) English, German, French, Swedish (higher-resource, multilingual scenario). For each language, we randomly sample up to 500 texts related to the topics of interest plus an equal amount of non-topical texts for training[9] in order to address the high imbalance of the dataset.

3.4 Empirical analyses

3.4.1 Analysis 1: Machine translation and multilingual representations

The first analysis investigates advantages and disadvantages of different combinations of algorithms and MT in terms of multilingual performance and usability. Figure 3.2 displays the average F1 macro performance for all 21 possible approaches in the high-resource scenario and 13 possible approaches in the low-resource scenario. The F1 macro values were calculated by giving each language the same weight and are an average over three random samples. The error bars display the standard deviation of F1 macro across all languages. For example, a value of

[9]Note that this kind of sampling would be more complicated in practical scenarios and approaches like active learning would be necessary to handle highly imbalanced datasets. As the focus of this paper is multilingualism and the language barrier, we leave improved sampling approaches to future work.

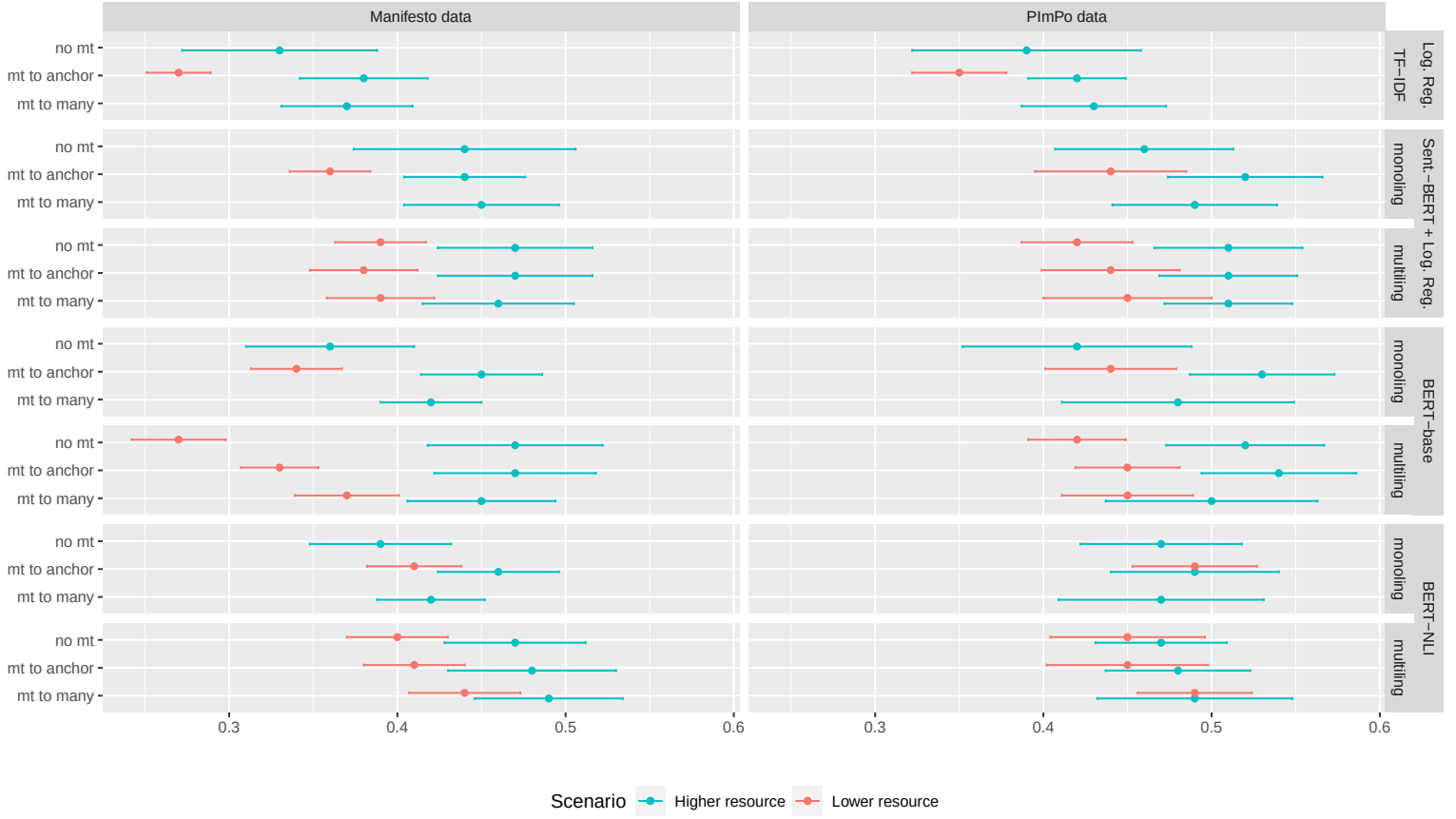
0.03 indicates that the performance for all languages fluctuates around +/-0.03 above or below the mean performance. We elaborate on our findings below.

Logistic Regression with TF-IDF vs. (m)BERT. All BERT variants clearly outperform the Logistic Regression with TF-IDF vectorization in the low-resource and higher-resource scenarios.[10] This is not surprising, as BERT can build upon prior 'knowledge' stored in their parameters, enabling them to learn new tasks faster.

Different (m)BERT variants. First, we find that the best performing BERT is the NLI variant. The only exception is PImPo in the higher-resource scenario. This is in line with prior research on English-only data, which shows that the prior 'task knowledge' of BERT-NLI enables it to learn new tasks faster than other BERT variants, while other variants become better as more data becomes available (Laurer et al., 2023). Our results show that this also holds for multilingual scenarios. Second, a surprising finding is that the Logistic Regression trained on the input from a (non-finetuned) Sentence-BERT performs similar to the fine-tuned BERT-base and performs more consistently well on the Manifesto Corpus. This seems surprising at first glance. One could assume that fine-tuning all parameters in BERT-base would enable it to specialise more in the specific task, while only fine-tuning a Logistic Regression on input from Sentence-BERT would not allow for sufficient adaptation to the task. There are probably two reasons why the results show the contrary: We used recommended hyperparameters for BERT-base (and -NLI) based on Laurer et al. (2023a) to reduce computational costs and we assume that the performance of these variants could have been slightly increased by a hyperparameter search. Based on the experience by Laurer et al. (2023a) the difference would probably not be large – BERT-NLI still performs best, although no hyperparameter search was conducted. Second, there is an established stream of research in the NLP literature showing that fine-tuning only a few new parameters on top of BERT, while keeping the main body untrained, can produce similar results as fine-tuning the entire BERT model (Houlsby et al., 2019; Pfeiffer et al., 2020). This is essentially what we are doing when training a Logistic Regression on the output from Sentence-BERT. This finding, also in a multilingual setting, is interesting, as it produces

[10] An interesting exception is BERT-base without MT and data augmentation. We assume its low performance is due to suboptimal hyperparameters. We did not conduct a hyperparameter search for fine-tuning the BERT models and, especially when little data is available, this can sometimes lead to model failures.

Figure 3.2: Average F1 macro performance for all possible approaches on two datasets in two scenarios



competitive results with a much lower computational budget.

Machine translation for data augmentation.[11] MT for data augmentation tends to improve performance with the 'mt to anchor' strategy, but the extensive augmentation with the 'mt to many' strategy is less beneficial. The classical Regression benefits from the 'mt to anchor' translation, as it enables the training of a single model with a larger quantity and variety of texts for the TF-IDF vectorizer. This performance cannot be consistently improved by the 'mt to many' augmentation, as this creates data in multiple languages and requires the training of different models for each language, which is impractical. Moreover, Sentence-BERT + Logistic Regression benefits from data augmentation in some scenarios, but not in others. This indicates that it is not clearly beneficial to introduce slight variations to the sentence embeddings via augmentation.

When fine-tuning BERT-base and BERT-NLI, data augmentation with MT does improve performance in the low-resource scenario, but the benefit is less clear in the high-resource scenario. When only very little data is available, mBERT benefits from the higher quantity and variety of data introduced through augmentation. It can ingest this multilingual augmented data, while the same monolingual variant cannot. At the same time, in the higher-resource scenario with texts from seven languages, the 'mt to many' scenario can even hurt performance. This is probably because too many (noisy) variations of the same texts are introduced (500 texts for 7 original languages multiplied by 7 translations in all other languages). Moreover, we assume that our standard hyperparameters are not ideally suited for all scenarios ranging from only 500 up to 24,500 training texts. For monolingual BERT, the best strategy seems to be translation to the anchor language. This enables the training of a single model on texts in one language, while both the no-MT and 'mt to many' strategy requires the training of different models for each language, which decreases average performance. Note that for monolingual models, the metrics in figure 3.2 are the average performance of multiple different models for each language.

English vs. Multilingual BERT. Overall, mBERT variants perform slightly better than monolingual (English) BERT trained on

[11]The exact train and test data for each scenario are defined in more detail in appendix D, tables 8 and 9. In figure 2 in the main text, 'mt to anchor' refers to 'one2anchor' in the low-resource scenario and 'many2anchor' in the high-resource scenario. 'mt to many' refers to 'one2many' in the low-resource scenario and 'many2many' in the high-resource scenario.

translated texts. This is probably influenced by multiple factors: First, mBERT can ingest all (multilingual) texts at the same time. If data in multiple languages is available (either in the higher-resource scenario, or through MT augmentation), a single BERT model can be trained on all languages. With monolingual BERT, a different model needs to be trained on each language separately with less data, which hurts performance. Moreover, it is possible that mBERT is better at identifying semantic patterns that are specific to individual languages, while these nuances are lost through machine translation and English-only transformers. We do, however, not have sufficient empirical evidence to substantiate this assumption. Third, for the 'monolingual' approaches with the no-MT or 'mt to many' strategy for languages other than English we were not able to use monolingual BERT variants, as they do not exist for all languages we analysed. For those that do exist (e.g. German and French), the performance would be very hard to compare due to very diverse design decisions of the creators of specific monolingual models (different training data, model architectures, hyperparameters etc.). We therefore only used mBERT in these cases with the respective monolingual input. These results therefore have to be interpreted with this caveat in mind. The lack of high-quality BERT variants in languages other than English is a relevant limitation in deep learning.

Variance of performance across languages. The paragraphs above only analysed aggregate performance for all languages at the same time. It is, however, also important to understand if certain approaches perform particularly well on specific languages, while failing on other languages. To better understand how (un)equally different approaches perform on different languages, we therefore also calculated performance on each language separately and report the standard deviation of performance across languages (indicated by the error bars in figure 3.2, see also tables 12 & 13 in appendix D3). The standard deviation allows us to report a single metric, while tables with disaggregated performance per language would have been hard to read and interpret. In the low-resource scenario, the standard deviation for different approaches is around 0.03 to 0.02 for Manifesto and 0.03 to 0.05 for PImPo. The difference between approaches is relatively small, but it is noticeable that high-performance algorithms (e.g. BERT-NLI) perform less equally across languages, while less performant approaches perform more equally (and badly) across languages. This indicates that performance increases can stem more from improvements on specific languages, although the

differences are not large. Surprisingly, in the higher-resource scenario, the standard deviation is higher overall (between 0.03 to 0.07). It would have been reasonable to assume that training data from all languages also leads to more equal performance on all languages. To the contrary, adding more data from more languages makes the classifiers' performance less equal across languages. It seems like adding more data from more languages increases the classifiers' performance on some languages more strongly while performance on other languages lags behind. Regarding the different combinations of MT and classification algorithms, we do not identify clear pattern for which approach systematically performs more/less equally across languages. In sum, we draw the following preliminary conclusions in terms of classification performance and usability based on our first analysis:

- BERT-NLI performs best, especially in the low-resource scenario, but requires additional experience to implement;
- A Logistic Regression trained on sentence embeddings by Sentence-BERT performs surprisingly well, while being computationally efficient and requiring less deep learning experience to implement;
- Multilingual models perform slightly better than monolingual models. In terms of usability, changing between monolingual or multilingual BERT effectively requires only changing one line of code and mBERT can avoid the need for potentially resource intensive MT;
- Data augmentation with MT improves performance in the low-resource scenario, but can hurt performance in the higher-resource scenario where higher quantity and variety of data is already available. Moreover, especially the 'mt to many' strategy requires additional time for implementation and compute;
- Surprisingly, performance across languages is less equal in the higher-resource scenario than in the low-resource scenario, indicating that some languages benefit more strongly from more data from other languages, while others benefit less.

3.4.2 Analysis 2: Validity test on political stances towards immigration

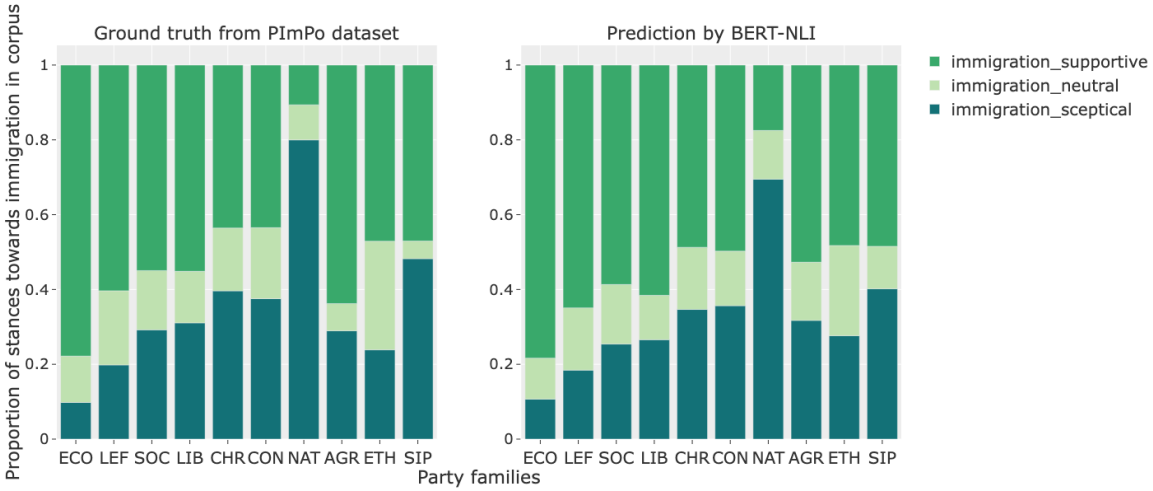
The second analysis goes beyond standard performance metrics and tests the validity of different approaches on parties' stance towards immigration. Regarding face-validity, we expect that left-wing parties are more supportive of migration, and the further to the right a party is situated, the more sceptical it is of migration (Zobel & Lehmann, 2018). A valid classifier should therefore produce this theoretically expected distribution. Regarding convergent validity, the PImPo dataset itself provides an external measurement of party positions along the left-right scale. A valid classifier should produce a prediction of party positions that correlates with the human assessment along the left-right scale in the full (held-out) PImPo corpus.[12]

Figure 3.3 displays the results from our best-performing model (BERT-NLI) trained on a sample of 1,674 English and German texts[13]. The figure provides a visualisation of the classifier's output to enable an assessment of face-validity (right sub-plot) and the output's correlation with the human ground truth (left sub-plot). Party families are ordered from left to right on the x-axis. The stacked bar-charts show the proportions of quasi-sentences that are supportive / neutral / sceptical towards immigration. For example, when Green parties ('ECO') mention immigration, 78% of mentions are supportive of immigration, 10% are sceptical, 12% are neutral according to crowd coders. The algorithm predicted a distribution of 78% supportive, 11% sceptical, 11% neutral. The proportion of supportive sentences linearly decreases from left to right, with nationalist parties 'NAT' being particularly sceptical. Note that agrarian ('AGR'), regional ('ETH') and special interest ('SIP') parties are special cases that do not fit the classical left-right scale. The model's output shows a linear increase of skepticism towards immigration from left to right as theoretically expected (face-validity) and it is strongly and significantly correlated with the human assessment (0.92 average

[12]Note that the classifier is trained on a small sample from the PImPo corpus, which makes this validation test not entirely external.

[13]To sample our training data, we sample up to 500 texts with stances and an equal number of texts that do not contain a stance towards the topic for each language. For example, this leads to a training dataset of 674 texts in English: 337 texts with stances and 337 without, as only 337 texts with stances are available in English. As more data is available in German, for example, we sample 500 texts with stances and 500 without. This leads to 674 texts for the English-only scenario and 1,674 for the English and German scenario. For more details, see appendix A2.

Figure 3.3: Comparison of true and predicted distribution of stances towards immigration by party families



correlation, convergent validity).

Table 3.1 provides details comparing the different approaches. The crucial additional metric is the average Pearson correlation coefficient. We calculated the correlation between the ground-truth proportion of sentences per party family being supportive of migration and the predicted proportion of sentences per party family being supportive of migration. We then calculate the same for sentences being skeptical of migration and neutral towards migration, resulting in three correlations and three p-values. The correlation and p-value columns are the average of these three respective values. The table is grouped by algorithm and ordered from highest to lowest correlation. As an additional test, we also added a sample extrapolation baseline. For the 'extrapolation-sample' row in the table, we calculated how well the training data sample alone can be a predictor for the data distribution in the (remaining) full corpus. Note that there is a relevant difference between correctly predicting overall proportions (approximated with the correlation coefficient) and correctly predicting individual texts. We therefore also provide the accuracy and F1 macro metrics as indicators for performance on individual texts. For most substantive analyses, the overall proportions in a corpus are arguably the most important measurement (Zobel & Lehmann, 2018).

Table 3.1: Correlation of algorithm predictions with human ground truth, stances by party family

algorithm	language representation	algorithm size	training languages	average correlation	average p-value	F1 macro	accuracy
BERT-NLI	multi	base	en-de-sv-fr	0.8	0.021	0.48	0.9
	multi	base	en-de	0.79	0.049	0.51	0.9
	multi	base	en	0.62	0.231	0.46	0.88
	en	large	en-de-sv-fr	0.84	0.022	0.51	0.91
	en	large	en-de	0.92	0.002	0.53	0.92
	en	large	en	0.92	0.001	0.53	0.91
	en	base	en-de-sv-fr	0.89	0.003	0.5	0.91
	en	base	en-de	0.78	0.062	0.52	0.91
	en	base	en	0.78	0.061	0.51	0.9
BERT-base/ -large	multi	base	en-de-sv-fr	0.5	0.138	0.44	0.89
	multi	base	en-de	0.66	0.253	0.43	0.85
	multi	base	en	0.51	0.16	0.45	0.88
	en	large	en-de-sv-fr	0.7	0.208	0.53	0.93
	en	large	en-de	0.74	0.144	0.52	0.91
	en	large	en	0.68	0.122	0.49	0.89
	en	base	en-de-sv-fr	0.64	0.323	0.47	0.89
	en	base	en-de	0.9	0.003	0.5	0.91
	en	base	en	0.17	0.154	0.33	0.72
Sent.-BERT + Log. Reg.	multi	base	en-de-sv-fr	0.78	0.057	0.47	0.92
	multi	base	en-de	0.74	0.09	0.46	0.9
	multi	base	en	0.5	0.2	0.47	0.91
	en	base	en-de-sv-fr	0.24	0.043	0.43	0.88
	en	base	en-de	0.28	0.077	0.46	0.9
	en	base	en	-0.08	0.368	0.38	0.84
extrapolation sample			en-de-sv-fr	0.73	0.112		
			en-de	0.27	0.034		
			en	0.26	0.072		

We make the following observations. Several algorithms achieve a surprisingly high correlation with the human ground truth of party families' stances towards immigration. BERT-NLI systematically correlates better with the human ground truth than the other BERT variants. Surprisingly, as opposed to the first analysis, the standard BERT-base (or -large) performs slightly better than Sentence-BERT, with one particularly well performing base-sized model[14]. Moreover, monolingual

[14]We assume that this high performance is linked to a lucky random initialisation of standard BERT's task-head. As we have also seen in the first analysis, standard BERT is more prone to sudden low or high performance when little data is available, as the task-head is randomly initialized from scratch for the new task. The other

English BERT tends to correlate better than mBERT. This is partly due to the fact that good mBERT variants are not available in large size. Base-sized English BERT also tends to correlate better for the NLI and standard variant, but the opposite is the case for Sentence-BERT. Unsurprisingly, larger BERT variants perform better overall, but the difference is not as large as one could have expected, as also base-sized BERT can achieve a good level of correlation.

Another surprise is that scenarios with more data from more languages do not necessarily lead to higher correlations. The two best performing algorithms were trained on only English or English and German texts and adding the two languages Swedish and French tends to reduce the correlation. This can have different reasons. First, when we add Swedish and French samples to the training data, we need to remove them from the full test corpus. Unfortunately, the PImPo corpus only contains very little data for certain languages (see appendix A). The evaluations for the scenarios with different amounts of training languages are therefore not directly comparable and changes in metrics can be due to changes in the test corpus. Second, more data might not always be better. The discourse in one language on migration might be very specific and confuse the algorithm or, on the contrary, the discourse in one language might be particularly representative for other languages and is sufficient to learn the task. Overall, it is encouraging to see that texts in only one language (English) from four countries (Australia, Ireland, New Zealand, United States) can be sufficient to achieve a significant average correlation of 0.92 with the human ground truth of party positions in eight other languages and ten other countries.

Moreover, an interesting finding is that the correlation with the positions of party families is only loosely linked to the machine learning performance metrics accuracy and F1 macro. One algorithm can have an F1 macro value of 0.53 and a very high correlation of 0.92, while another has the same F1 macro performance, but only a correlation of 0.70. This is even worse for accuracy, which is above 0.9 for many algorithms, given the high imbalance of the dataset, but the average correlation of a 'highly accurate' algorithm can be as low as 0.28. One relevant reason for the loose link between classification metrics and the correlations is that we used slightly different data to calculate them. The calculation of the classification metrics also needed to include the 'not related to immigration' class, which we excluded from the correlation analysis

BERT variants do not have this instability issue.

following Zobel and Lehmann (2018). This also indicates, however, that machine learning metrics designed to assess the ability of a classifier to accurately reproduce classification tasks are not necessarily a good indicator for how well a classifier can predict a substantively interesting data distribution like positions towards immigration by party family. As each substantive use-case has different requirements (different classes might have different importance, performance on specific subsets of the data might be more important) it is difficult to recommend a specific acceptability threshold for these metrics.

3.5 Discussion and conclusion

Our toolkit of accurate, practical and valid computational text analysis methods is still very much in development. We know surprisingly little about how to produce valid results that are practically useful for comparative social science research (Baden et al., 2022). In this paper, we test several recent innovations from deep transfer learning to advance our computational toolkit for multilingual social science research. We demonstrate that, based on these innovations, supervised classifiers can produce substantively meaningful output. BERT-NLI trained on only 674 or 1,674 texts in only one or two languages can validly predict political party families' stances towards immigration in eight other languages and ten other countries.

In two interlinked analyses, we asked: what are the advantages and disadvantages of different algorithms in multilingual settings in terms of performance and usability? What are the advantages and disadvantages of combining MT with multilingual BERT? And how can validity in multilingual text classification be assessed and how do different approaches impact validity? To answer these questions, we analysed two datasets with texts in 12 languages from 27 countries with 21 different computational approaches.

We compare each supervised approach based on its performance, usability and output validity. We find that BERT-NLI performs best, both in terms of classification performance and output validity. Its disadvantage is the required computational resources and knowledge for implementation, and its value is reduced when more than around 2,000 annotated texts are available. Combining Sentence-BERT with a Logistic Regression could be a computationally cheaper and simpler alternative. It performs surprisingly well on classification metrics, but surprisingly

badly in our validation test. The comparison of multilingual or English BERT led to mixed results. In the first analysis, mBERT tended to perform better, while English BERT variants were better in the second analysis. As there is no large difference between the two, we tentatively recommend using English BERT(-NLI) in multilingual settings with texts machine translated to English. Working with English texts enables more researchers to understand the data they are analysing; the quality of machine translation is constantly improving, with free, open-source MT models becoming available; and English BERT is available in a wider variety of sizes, tasks and higher quality. Our attempts to boost the potential of mBERT by combining it with data augmented with MT mostly led to improvements in low-resource settings. The additional effort necessary for the 'mt to many' augmentation strategy probably does not warrant the additional resources and risks of introducing noisy data. When it comes to validation, we demonstrate that, if applied well, supervised classifiers can produce valid and substantively meaningful output.

Based on these findings, we recommend using English BERT-NLI with machine translation to English if the required expertise is available in the research team and less than 2,000 annotated texts are available. Otherwise, we recommend the combination of Sentence-BERT and Logistic Regression.[15] For implementing the BERT-NLI approach, we recommend re-using the easy-to-use Google Colab code provided in (Laurer et al., 2023a), which also enabled the analyses in this paper.[16]

Our comparative analysis of multilingual supervised machine learning is, however, subject to several limitations. First, future research should dive deeper into qualitative analyses of additional reasons for high or low performance. Second, our second analysis is based on a balanced training dataset, which would be difficult to create in practice. More advanced strategies for sampling training data should be further investigated in future research. Third, we did not explicitly explore questions of (political) bias. While our findings are encouraging, as our best classifiers represented stances towards immigration well independently of party families, more research specifically on political bias in language models is necessary. Fourth, while we demonstrate the face- and convergent validity of classifiers using one prominent multilingual example, more

[15]The Hugging Face SetFit library provides an easy-to-use and more advanced implementation of this approach: `https://github.com/huggingface/setfit`

[16]`https://github.com/MoritzLaurer/less-annotating-with-bert-nli`

diverse tests on more datasets are necessary to fully understand the conditions under which supervised machine learning produces valid and substantively meaningful outputs.

Can we lower the language barrier for comparative research with supervised machine learning? Yes, our analysis shows empirically that both deep transfer learning and machine translation enable us to produce meaningful computational results in multilingual text analysis research. By using state-of-the-art methods, we can leverage around a thousand texts in one or two languages to produce results that apply to many other languages and countries. We hope that future research can build upon our findings and apply the most promising methods to a more diverse set of tasks, domains and research questions.

Appendix

The appendix is available online via the published version of the paper: *Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023). 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*

Chapter 4

On Measurement Validity and Language Models

Increasing Validity and Decreasing Bias with Instructions

Abstract. *Language models like BERT or GPT are becoming increasingly popular measurement tools, but are the measurements they produce valid? Literature suggests that there is still a relevant gap between the ambitions of computational text analysis methods and the validity of their outputs. One prominent threat to validity are hidden biases in the training data, where models learn group-specific language patterns instead of the concept researchers want to measure. This paper investigates to what extent these biases impact the validity of measurements created with language models. We conduct a comparative analysis across nine group types in four datasets with three types of classification models, focusing on the robustness of models against biases and on the validity of their outputs. While we find that all types of models learn biases, the effects on validity 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 produce more valid measurements across different groups. An instruction-based model (BERT-NLI) sees its average test-set performance decrease by only 0.4% F1 macro when trained on biased data and its error probability on groups it has not seen during training increases only by 0.8%.*

Paper published as: *Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2024). On Measurement Validity and Language Models: Increasing Validity and Decreasing Bias with Instructions. Communication Methods and Measures, 1–17. https://doi.org/10.1080/19312458.2024.2378690*

4.1 Introduction

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

In recent years language models have become an increasingly popular and accurate measurement tool, but do these models also produce more valid measurements? Computational social scientists have warned about the challenges of validity and computational text analysis methods for a long time (Grimmer & Stewart, 2013) and recent evidence suggests that there is still a relevant gap between the ambitions of our tools and the validity of their outputs (Baden et al., 2022). In this paper we try to narrow this gap, by investigating the relationship between new language models and validity.

As validity is a very broad and ambiguous term, we specifically focus on *measurement validity* as defined by Adcock and Collier (2001) and how it is impacted by biases in machine learning training data. This narrow focus is inspired by the natural language processing (NLP) fairness literature, which argues that language models like BERT or GPT behave like "stochastic parrots" that reproduce (spurious) patterns from their training data instead of truly understanding the concepts they are intended to measure (Bender et al., 2021). We follow a group-based definition of bias, where a model is considered biased if it performs unequally across social groups. A source of bias can be unequal representation of groups in the training data (biased data), where certain group-specific language patterns are spuriously correlated with the concepts we want to measure (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021). This group-based bias is particularly problematic for social science research, as social scientists often need to validly measure complex concepts across different types of social groups such as countries, milieus or languages, while working with real-world training data that is never perfectly representative. While there is ample research in the NLP literature about bias, there is little empirical research by social scientists on how bias impacts measurement validity and how its impact

on validity can be mitigated.

Given this gap in the social science literature (Baden et al., 2022), this paper empirically investigates the research questions: To what extent do biases in fine-tuning data impact validity in supervised machine learning for social science tasks (RQ1)? How robust are different supervised machine learning approaches against biases in fine-tuning data (RQ2)? Do meaningful instructions for language models reduce bias and increase validity (RQ3)?

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 decrease bias and 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 spurious group-specific language patterns from the fine-tuning data (section 2).

We test our assumptions empirically, by analysing the interaction between bias and test-set performance of different models (section 3 and 4). We fine-tune three types of text classifiers on texts from four datasets and nine different group types 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, e.g. one country) reduces performance on representative test sets (from all groups).[2] 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 more likely to produce 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 groups it has not seen during training increases from 26% to 31.1%. An instruction-based

[1]The full reproduction code and data is available at `https://github.com/MoritzLaurer/language-model-bias-validity`

[2]By "group type", we refer to a group category in the data like "country", "political party" or "decade". By "group", we refer to a member of a group type. France or Germany are groups within the group type "country".

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 groups it has not seen during training increases only from 18.5% to 19.3%.

4.2 Measurement validity and bias in computational text analyses

4.2.1 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).[3] For the purpose of this paper, we use the term "measurement validity" as the main type of validity based on Adcock and Collier (2001). Their 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.

Simply put, a measurement is valid, when it actually measures what the researcher wants it to measure (Adcock & Collier, 2001). To systematize this definition, it is helpful to make the process of measurement explicit using the example of a supervised machine learning (SML) 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

[3]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 & Collier, 2001, p. 536-537). The causal inference literature emphasizes the terms internal and external validity (Cook & Campbell 1979, based on Adcock & Collier, 2001, p. 529). The content analysis literature discusses at least 12 different types of validity (Krippendorff, 2018). The political science text-as-data literature uses yet another set of complementary, but also different terms (Benoit, 2020; Grimmer, Roberts, & Stewart, 2022; Grimmer & Stewart, 2013). Only very few interdisciplinary efforts for harmonizing terminologies from a computational perspective exist (Jacobs & Wallach, 2021).

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 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 could then annotate ('score') several hundred sentences based on the pre-defined classification scheme (level 4). If the full textual corpus is too large for manual annotation, the researchers could then train a supervised classifier to automatically classify ('score') the remaining sentences in the full corpus of thousands of party manifestos. The researchers then need to aggregate the classifier's predictions for individual sentences ('scores') into the indicator, for example by calculating the proportion of populist sentences relative to all sentences per year and per country.[4] 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.

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.[5] As there is a broad literature on measurement validity, this paper only focuses on two specific aspects linked to measurements derived from supervised machine learning models: First, we investigate how group-based biases in the training data can impact the scoring error at level 4. Second, we hypothesize

[4]Note that Adcock and Collier (2001) add a "Refining Indicators" step here, where the classification scheme can be refined during initial iterations over the data.

[5]Note that we follow (Adcock & Collier, 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.

Figure 4.1: Main steps for creating a valid measurement

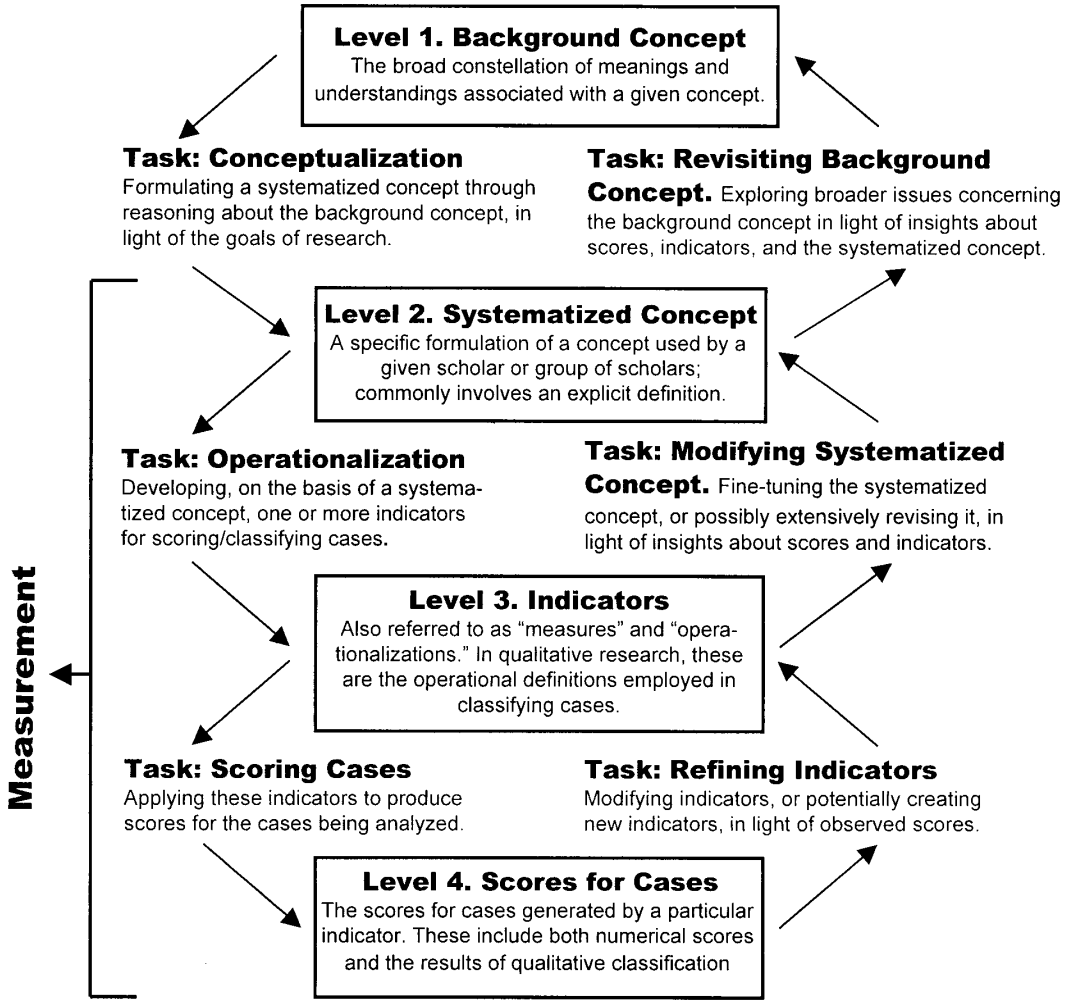


Figure from Adcock and Collier (2001, p. 531)

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 validity. As the main type of validation in this paper, we use test-set validation combined with additional statistical tests.

4.2.2 Robustness against group-specific patterns as a precondition for validity

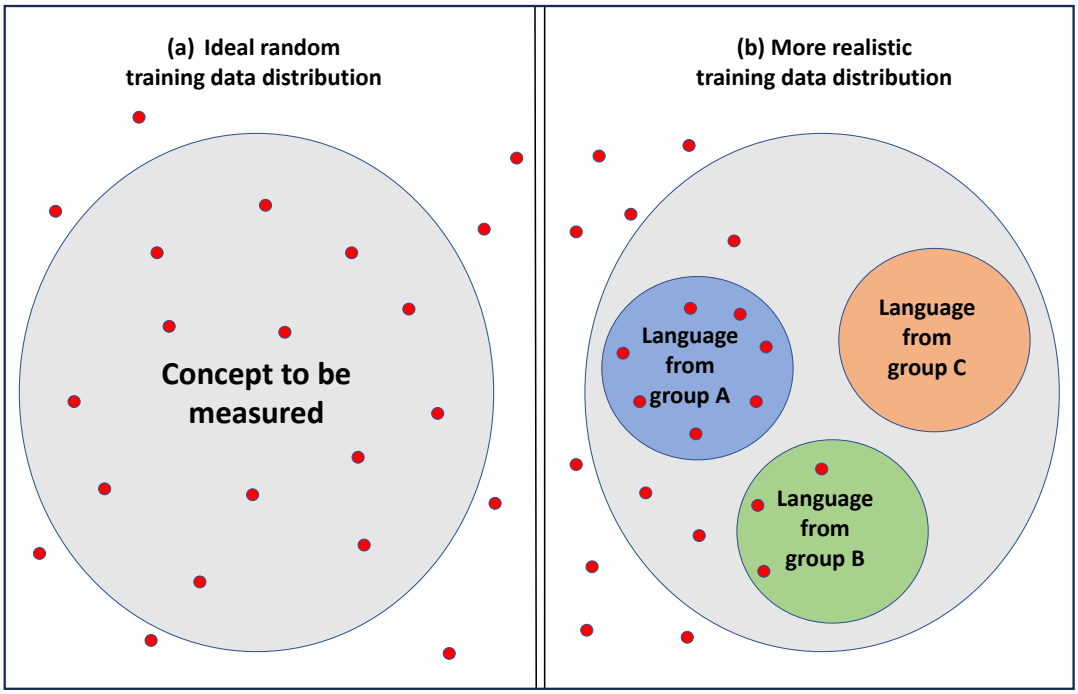
Why are group-specific patterns relevant for validity? In the standard machine learning process, the only source of information for learning the concepts of interest and the related scoring task is the training data. A classical logistic regression or support vector machine does not have any prior knowledge 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 if they are trained following the standard pre-train-fine-tune paradigm (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 (Laurer et al., 2023a). Both a logistic regression and BERT-base are designed to find any pattern in the training data that help them reduce their error on their training data. They can be seen as "stochastic parrots" (Bender et al., 2021). These models can only fully learn a concept of interest, if the training data comprehensively represents all ways of expressing the concept of interest.[6] This ideal scenario is quite unrealistic (figure 2, left). A more realistic scenario is illustrated in the right part of figure 2. In practice, researchers will have access to imbalanced data from a few social groups (e.g. specific countries) 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 (distribution shift, see Ruder (2019)). [7] This directly impacts measurement validity: measurements are only valid if they measure what we think they measure.

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, He, Zou, Tao, & Hu, 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

[6] A BERT-base model can only partly mitigate this issue through its prior language knowledge of synonyms etc.

[7] 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, which further increases complexity.

Figure 4.2: 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 (a single class). Red dots represent individual training data points. Red dots outside of the grey area do not represent the concept of interest (e.g. another class).

('position bias'), overlapping terms for bi-text tasks ('overlap bias'), or specific writing styles that 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 (X. Wang, Wang, & Yang, 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) 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.

4.2.3 Increasing robustness and validity with instructions

How can this issue be addressed, where models learn spurious patterns and biases instead of the actual concept we want to measure? We argue that the standard training or fine-tuning procedure is at the core of these issues of bias and validity.[8] From a validity perspective, the information the model receives during standard 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: Imagine we want to measure eight types of emotions in a large corpus of a million social media posts. We recruit crowd workers from the internet to help us analyse these texts. To teach them about our task and concepts, we send them an email with a few hundred example texts that have been categorised into one of the eight emotions. In the texts we send, the category for each text is only indicated by a number in the title. We do not provide any explanation what the task and categories are

[8]Note that our argument applies both to training classical models like logistic regressions as well as standard fine-tuning of BERT models. We therefore use the terms 'standard fine-tuning' and 'training' synonymously for simplicity in this paper.

about. Only based on a few hundred texts and numbers, how many crowd workers would understand that we want to measure eight specific types of emotions from a specific psychological theory? While sending numbered texts without any explanations or definitions would obviously be a bad teaching process for human learners, this is effectively how the standard training procedure teaches a task to a model. In the case of a logistic regression, the learner 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 we are interested in (a literate but ill-instructed parrot). It seems unsurprising that these learners can produce invalid measurements.

This is why, in practice, a key additional input for teaching a task to human annotators is a codebook with clear definitions of the concept of interest. Throughout the annotation process, a diligent research assistant or crowd worker can refer back 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).[9] 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, 2023b), to prompting masked language models like BERT (Schick & Schütze, 2021a), to combining universal tasks like Natural Language Inference (NLI) with BERT models (Laurer et al., 2023a; Yin et al., 2019). 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, Maini, Kolter, Lipton, & Pruthi, 2023). Only very few papers have applied instruction-based models in the social sciences (Argyle et al., 2023; Laurer et al., 2023a; Laurer, Van Atteveldt, Casas, & Welbers, 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:

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

(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 robustness of instruction-based models against group-specific language patterns and the consequences for validity. 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 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. Instruction-based models, on the other hand, re-use all their parameters for new tasks and do not need to reinitialize parameters. Raman et al. (2023) argue that it is this algorithmic difference that makes instruction-based models more robust against spurious patterns in training data instead of the semantics of instructions. We test both possible explanations empirically below.

4.3 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; and sufficient quantity of data for training and testing across data splits for different groups.

We compare the following classification models:

- Logistic regression as a representative for classical machine learning approaches (illiterate and ill-instructed parrot);
- DeBERTa-v3-base as a representative of standard transfer learning

Table 4.1: Overview of datasets used in the study

Dataset	Task & Concepts	Text domain	Groups	Data size
PImPo (Zobel & Lehmann, 2018)	Identify stances towards Immigration/Integration (4 classes: supportive, sceptical, 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 (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 (CAP-SotU merged with CAP-Court) (Project, 2014)	Identify five topics ('Domestic Commerce', 'Law and Crime', 'Civil Rights', 'Labor', 'Government Operations')	US presidential speeches & US court rulings	2 domains (speeches / legal text)[1]	Train: 7708 Test: 1928

[1] These two types of domains are not social groups, but we include them to test a scenario where the language between two groups of text is particularly different. This type of hard domain shift is a common challenge.

approaches (literate but ill-instructed parrot). DeBERTa-v3 is a BERT variant that strongly outperforms the original BERT model (P. He et al., 2021). We refer to it as 'BERT-base' in the remainder of the text for simplicity;

- DeBERTa-v3-base fine-tuned for the universal Natural Language Inference task ('BERT-NLI') as a representative for instruction-based approaches (literate and instructed parrot). BERT-NLI can ingest instructions in the form of "class hypotheses". For a stance detection task, for example, the class hypotheses could be "This text is positive towards the military" and "This text is negative towards the military", which are fed into the model as a third input. See Laurer et al. (2023a) for a more detailed

explanation. The exact instructions used for each dataset are provided in appendix B. 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.[10]

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

1. To what extent do biases in fine-tuning data impact validity in supervised machine learning for social science tasks?

We approximate the impact of group-based biases in the fine-tuning data on validity in two steps: First, each classification model is trained on texts sampled from only one group (e.g. only one country, "biased condition"). Second each model is also trained on texts randomly sampled across all groups ("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 groups (see Appendix D for the data distributions of all datasets). 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, making it harder to extrapolate to other groups during testing. We conduct test-set validation with the standard classification metric F1 macro.[11] 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 during training.

This study design simulates extreme situations of bias in the training data by only sampling from one group, while in practice researchers will often have access to data from more groups. This setup is designed to give us a clear idea of the impact of bias from group-specific language patterns. Another reason for this choice is to enable comparability across

[10]The exact BERT-NLI model used is available at `https://huggingface.co/MoritzLaurer/deberta-v3-base-zeroshot-v1`

[11]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.

different datasets, as two of our datasets only have group types with maximum two groups (see table 1).[12] Also note that this paper focusses on analysing the difference in robustness of different types of classifiers against biases across datasets and the role of instructions, instead of the reduction of bias in a specific case-study.[13]

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 groups have very little data for some classes and with 500 texts there are still enough groups that have enough texts for minority classes.

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-based biases. Note that we do not conduct hyperparameter searches for our experiments and use the recommended parameter values determined by an extensive hyperparameter search by Laurer et al. (2023a), as a hyperparameter search across this wide range of configurations (models, datasets, groups, bias) runs would be prohibitively expensive.

We acknowledge 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

[12]Only for the CoronaNet country group type, we sample from three groups 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.

[13]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.

each dataset. Given this assumption, test-set validation with classification metrics is a good validation procedure that can be implemented comparatively across multiple datasets.[14]

2. How robust are different supervised machine learning approaches against biases 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.[15] 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 or 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 a row in our test data frame contains a prediction on a text from group A by a classifier that was also (only) trained on group A, we flag it as a "biased row" in our tabular data frame. 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 classifier type and bias on the error in the test data. The interaction between classifier type and biased rows results in odds ratios which let us draw conclusions on the degree of bias of different classifier types.

[14]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 design across datasets.

[15]More concretely, the underlying data are 922224 observations (i.e. predictions on test-set texts) from the 3 types of models on 4 datasets, 9 types of groups across 6 random seeds from the biased condition form the first analysis.

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

Lastly, we test the hypothesis that meaningful instructions provided to language models can function similarly to instructions provided to crowd workers, reducing group-based biases and increasing validity. For BERT-NLI, the researcher manually formulates short instructions that verbalize each class based on the codebook that guided human annotators.[16] 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.

4.4 Empirical analyses

RQ1: To what extent do biases in fine-tuning data impact validity in supervised machine learning for social science tasks?

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 across all groups, 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 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).

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

Figure 4.3: Test set validation and bias penalty in aggregate

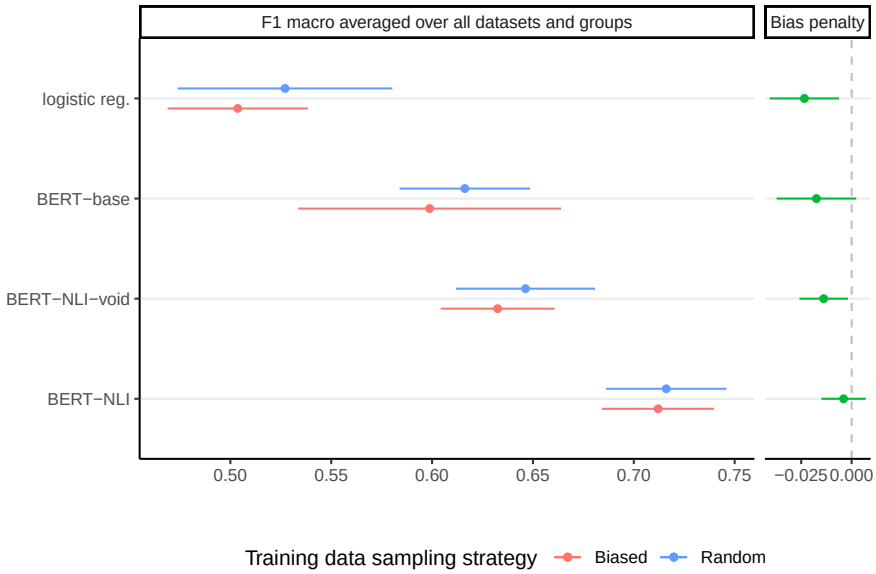
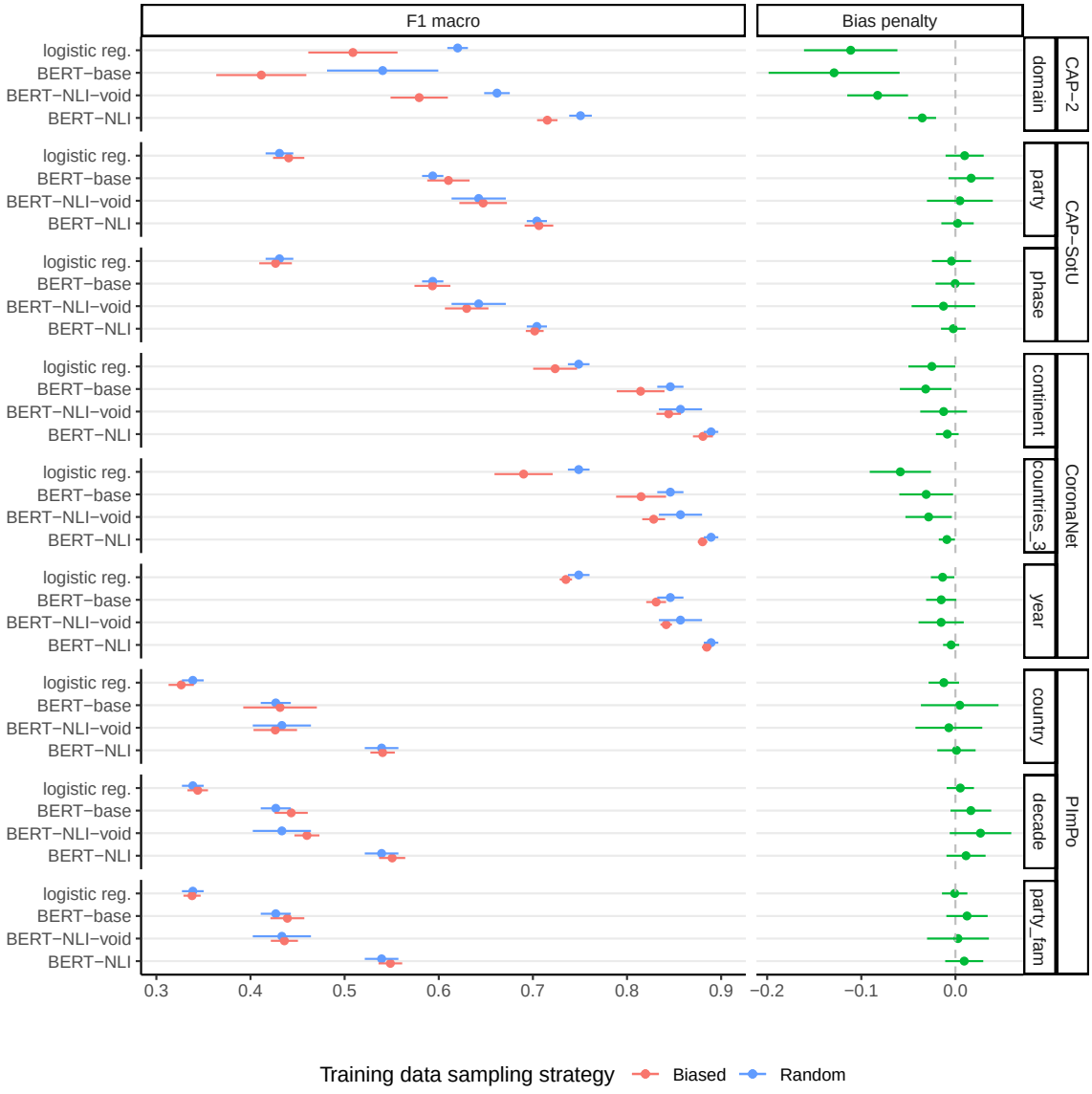


Figure 4 provides a more nuanced, disaggregated picture per group and dataset. The error bars display the standard deviation across six random runs with different random seeds and groups. For some datasets and groups, the bias penalty is very small while it is larger for others. We also notice that for a few combinations (PImPo and CAP-SotU by 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 groups in the test data. As the test data is randomly sampled from real-world datasets, it contains more data from some groups 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 groups 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 groups sampled for training can therefore also constitute the majority groups 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 4.4: Test set validation and bias penalty by group types



RQ2: How robust are different supervised machine learning approaches against biases in fine-tuning data?

We used a binomial mixed-effect regression model to test the effect of group-based biases, and compare this effect across classifiers. For this analysis we look more closely at the errors in the biased training runs (represented by the red dots in the preceding figures). Our model shows how much more likely it is for a classifier to make an error if the test text comes from a group that the classifier has not seen during training. We argue that the strength of this effect serves as a measure of group-based bias, because decreased error on groups seen during training indicates reliance on group-specific language patterns. We fit this model to data from all classifiers, and used interaction effects to test and compare this effect across classifiers. For each classifier we pooled runs from all six random seeds and all datasets, and included random intercepts for every combination. We thereby measure the average effect of group-based biases across different classifier types.

Figure 5 presents the odds ratios for how less likely an error is on data from groups seen during training (right), as well as the corresponding error probabilities (left). The red dots represent the classifiers' probability of making an error on test data that comes from the same group as the classifier has been trained on, while the blue dot represents the error probability on data from groups the classifier has not seen during training. The full regression tables and the figure disaggregated by group are available in appendix A.

For the logistic classifier, we see that the odds of making an error decrease by a factor of 0.84 (SE: 0.01, $p < 0.001$, 95% CI [0.82, 0.85]) when the text comes from the same group the classifier has seen during training. This is a clear indication of bias. In other words, the probability of making an error on text from groups it has not yet seen during training is 40%, while the probability of making an error on texts from groups 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 (SE: 0.01, $p < 0.001$, 95% CI [0.92, 0.97]), 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 most by a factor of 0.78 (SE: 0.01, $p < 0.001$, 95% CI [0.76, 0.80]) on texts from groups it has already seen during training (26% probability of error) compared to groups it has

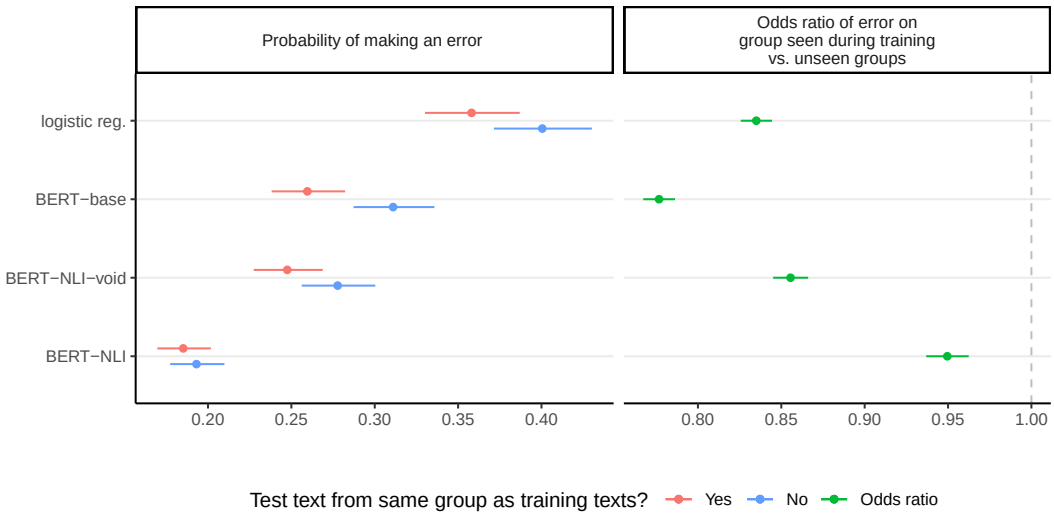
not seen during training (31.1% probability of error). This indicates that BERT-base makes less mistakes than the logistic classifier overall, but a part of its performance advantage comes from learning more group-specific language patterns (overfitting).

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 for language models reduce bias and increase validity?

The results discussed above show that BERT-NLI is both less biased and performs better in terms of test set validation 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

Figure 4.5: Error analysis and bias



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 (partially contradicting Raman et al. (2023)). 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 4). 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 according to the instruction text.

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.

4.5 Limitations and discussion

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 (next-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 the same text classification performance as BERT-like models (Schick & Schütze, 2021a; H. 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 (P. He et al., 2021, p. 8). "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. Moreover, for both generative models and BERT-NLI, varying the formulation of the instructions (prompts) is an important avenue for optimization (essentially a hyperparameter), which we did not investigate in this study due to computational limitations.

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; X. Wang et al., 2022) as well as in NLI data (Gururangan et al., 2018) from the NLP fairness community. Our paper focuses on group-based biases in the fine-tuning data that are relevant for comparative social scientists and 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 comparative validation and bias analyses across a wider array of datasets. Several additional methods like feature importance analysis or manual error analysis exist and are particularly suitable for understanding individual datasets and models more deeply. 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 multiple datasets, and we complement this analysis with the binomial mixed-effect regression as an additional statistical test.

4.6 Conclusion

This paper investigates the effect of group-based biases in machine learning training data on measurement validity. We show that all types of classifiers learn group-based 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 groups which 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 groups 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 groups 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 groups (especially for shifts from legal to speech language and partly between countries) and smaller for other groups (political parties or time periods).

We argue that the high level of robustness against bias and test-set validity of instruction-based BERT-NLI is due to two important characteristics. First, on an algorithmic level, instruction-based models do not need to delete and randomly re-initialize task-specific 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.

What do our results mean for researchers in practice? First, for research projects that use text classification for measurement, we recommend using the instruction-based BERT-NLI models, especially when little training data is available. We assume that the bias penalty decreases as more and more balanced data becomes available and standard BERT-base models become a more viable option (based on Laurer et al., 2023a). For researchers who are new to these methods, we recommend following a workshop by the first author which also includes copy-pasteable Juypter notebooks with reusable training code and a 4 hour video with additional explanations.[17] Second, we recommend paying attention to group-imbalance in addition to class-imbalance both in the training and test data. If a high quality test dataset exists, researchers can gain confidence in their models by calculating disaggregated metrics for all substantially relevant groups to identify potential model biases. Researchers should iteratively improve their data if they

[17] `https://github.com/MoritzLaurer/summer-school-transformers-2023/tree/main`

identify issues.

Lastly, 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.

Appendix

The appendix is available online at `https://osf.io/2t4cd`.

Chapter 5

Building Efficient Universal Classifiers with Natural Language Inference

Abstract. *Generative Large Language Models (LLMs) have become the mainstream choice for fewshot and zeroshot learning thanks to the universality of text generation. Many users, however, do not need the broad capabilities of generative LLMs when they only want to automate a classification task. Smaller BERT-like models can also learn universal tasks, which allow them to do any text classification task without requiring fine-tuning (zeroshot classification) or to learn new tasks with only a few examples (fewshot), while being significantly more efficient than generative LLMs. This paper (1) explains how Natural Language Inference (NLI) can be used as a universal classification task that follows similar principles as instruction fine-tuning of generative LLMs, (2) provides a step-by-step guide with reusable Jupyter notebooks for building a universal classifier, and (3) shares the resulting universal classifier that is trained on 33 datasets with 389 diverse classes. Parts of the code we share has been used to train our older zeroshot classifiers that have been downloaded more than 65 million times via the 🤗 Hugging Face Hub as of March 2024. Our new classifier improves zeroshot performance by 9.4%.*

Preprint published as: *Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2023). Building Efficient Universal Classifiers with Natural Language Inference (arXiv:2312.17543). arXiv. https://doi.org/10.48550/arXiv.2312.17543*

5.1 Introduction

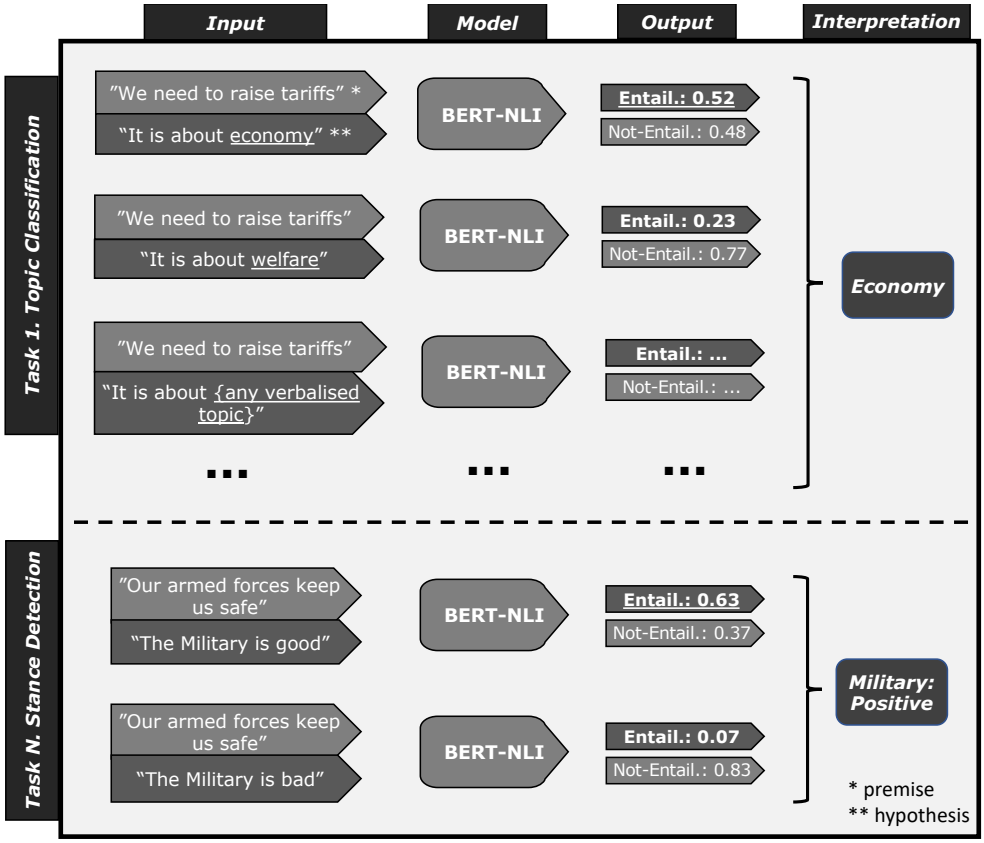
Over the past year, generative models have taken both academia and public attention by storm. The main appeal of text generation is that it is so universal, that almost any other text-related task can be reformulated as a text generation task (Radford et al., 2019; Raffel et al., 2020). Especially when text generators are massively scaled up and tuned on human instructions, they acquire impressive capabilities to generalise to new tasks without requiring task-specific fine-tuning (Chung et al., 2022; OpenAI, 2023b; Ouyang et al., 2022; Sanh et al., 2022; Touvron et al., 2023). Since the utility of these generative Large Language Models (LLMs) has become evident, large amounts of intellectual, financial and energy resources are being invested in improving and scaling generative LLMs.

Given that the resource requirements for training and deploying generative LLMs are prohibitive for many researchers and practitioners, this paper investigates other types of universal models, that make a different trade-off between resource requirements and universality. The literature has developed several other universal tasks that cannot solve generative tasks (summarization, translation etc.), but can solve *any* classification task with smaller size and performance competitive with generative LLMs (Schick & Schütze, 2021b; H. Xu et al., 2023).

Building better classifiers is particularly important in the social sciences, where classification is often used as a measurement tool. Social science researchers need models that are easy to use, while producing valid measurements across all classes and social groups of interest. Universal classifiers can be an important addition to this toolkit, as they can either circumvent the need for fine-tuning altogether (zeroshot classification) or they can be a base model for robust task-specific fine-tuning with less data (Laurer et al., 2023a; Laurer, Van Atteveldt, Casas, & Welbers, 2023c).

The principle of universal classifiers is similar to generative models: A model is trained on a universal task, and a form of instruction or prompt enable it to generalize to unseen classification tasks. While several efficient approaches to universal classification exist (Bragg, Cohan, Lo, & Beltagy, 2021; Ma, Yao, Lin, & Zhao, 2021; Schick & Schütze, 2021a; Sun, Zheng, Hao, & Qiu, 2022; Xia et al., 2022; H. Xu et al., 2023; Y. Yao et al., 2022), this paper focuses on guidance for one approach: Natural Language Inference. Several papers have used the universal NLI task for zero- and fewshot classification, but stopped short of mixing

Figure 5.1: Illustration of universal classification with BERT-NLI based on Laurer et al. (2023a)



NLI data with multiple other non-NLI datasets to build more universal classifiers (Laurer et al., 2023a; S. Wang et al., 2021; Yin et al., 2019, 2020).

The main contribution of this paper are: (1) easy-to-use universal classifiers trained on 5 NLI datasets and 28 non-NLI datasets with 389 diverse classes, improving zeroshot performance by 9.4% compared to NLI-only models; (2) a step-by-step guide with Juypter notebooks enabling users to train and adapt their own universal classifiers.

Figure 5.2: Example for using the resulting universal classifiers in the 🤗 zeroshot pipeline

```python
from transformers import pipeline

text = "Angela Merkel is a politician in Germany and leader of the CDU"
hypothesis_template = "This text is about {}"
classes_verbalized = ["politics", "economy", "entertainment", "environment"]

model_name = "MoritzLaurer/deberta-v3-large-zeroshot-v1.1-all-33"
classifier = pipeline("zero-shot-classification", model=model_name)
classifier(text, classes_verbalized, hypothesis_template=hypothesis_template)

# output: {'labels': ['politics', 'entertainment', 'economy', 'environment'],
    'scores': [0.99, 0.0, 0.0, 0.0]}
```

5.2 NLI as a universal task

The Natural Language Inference (NLI) task[1] is defined as recognising if the meaning of one text (the hypothesis) is entailed in another text (the premise). For example, the hypothesis "The EU is not trustworthy" is entailed in the premise "The EU has betrayed its partners during the negotiations on Sunday". To create NLI datasets, workers are presented with a text (the premise) and are tasked with writing a hypothesis that is either clearly true given the premise (entailment), clearly false given the premise (contradiction), or that might be true or false but is not clearly entailed or a contradiction (neutral). Several large scale NLI datasets with hundreds of thousands of unique hypothesis-premise pairs for these three classes have been created by crowd workers or language models (Bowman, Angeli, Potts, & Manning, 2015; Conneau et al., 2018; Liu, Swayamdipta, Smith, & Choi, 2022; Nie et al., 2020; Parrish et al., 2021; Williams et al., 2018). For simplicity and to increase universality, the task can be simplified into a binary entailment vs. not-entailment task by merging the 'neutral' and 'contradiction' labels (Yin, Radev, & Xiong, 2021).

This binary NLI task is universal, because any text classification task can be reformulated into this entailment vs. not-entailment decision through label verbalisation (see figure 5.1). Take topic classification as

[1] An older but more expressive name for the task is RTE, Recognising Textual Entailment (Dagan et al., 2006)

an example. The task could be to determine if the text "We need to raise tariffs" belongs to the topic "economy" or "welfare". From an NLI perspective, we can interpret the text "We need to raise tariffs" as the premise and verbalise the topic labels in two topic hypotheses: "This text is about economy" and "This text is about welfare". The classification task reformulated as an NLI task then consists of determining which of the two topic hypotheses is more entailed in the text of interest (premise). In different words: Which hypothesis is more consistent with the text of interest?

A model fine-tuned on NLI data (e.g. "BERT-NLI") can then be used to test any hypothesis formulated by a human against any text of interest (premise). For each individual hypothesis-premise pair, an NLI models will output a probability for entailment and not-entailment. To choose the most probable topic, we can select the hypothesis with the highest entailment score. Following the same procedure, any other text classification task can be reformulated as an NLI task, from stance detection, sentiment classification to factuality classification (see figure 5.1). Any class can be verbalised as a hypothesis (similar to the prompt of a generative LLM) and can then be tested against any text of interest.[2]

The main disadvantage of NLI for universal classification is that it requires a separate prediction for each of N class hypotheses, creating computational overhead for tasks with many classes. The main advantage is that identifying a new class only requires verbalising it as a hypothesis and passing it to an NLI model without the need of fine-tuning a new task-specific model from scratch (zeroshot classification). The most prominent implementation of this approach is probably the Hugging Face `ZeroShotClassificationPipeline` (see figure 5.2) which uses this NLI-based approach under the hood (Wolf et al., 2020).[3] The models created in the paper are designed to be directly compatible with this pipeline.

5.3 A guide to building a universal classifier

In this guide we explain how this type of universal classifier is built. Each step is accompanied by a Jupyter notebook available on GitHub

[2]Note that an NLI model will always only do one task (NLI) just like a GPT model can only predict the next token. These tasks are universal because any other specific task can be reformatted into these more general tasks.

[3]`https://huggingface.co/docs/transformers/v4.36.1/main_classes/pipelines`

that implements each step end-to-end.[4] The main steps are:

1. Dataset preprocessing and harmonization
2. Automatic data cleaning (optional)
3. Hypothesis formulation and formatting
4. Training and evaluation
5. Visualisation of results

Guidance for using the resulting model is provided in section 4.

5.3.1 Data selection, preprocessing and harmonization

We use two main types of data to train our universal classifier: Five NLI datasets and 28 other classification datasets.

`data-harmonization-nli.ipynb`

First, we use a set of established NLI datasets: MNLI (Williams et al., 2018), ANLI, FEVER-NLI (Nie et al., 2020), WANLI (Liu et al., 2022), Ling-NLI (Parrish et al., 2021).[5] Each dataset contains tens of thousands of unique hypothesis-premise pairs classified into one of the three classes "entailment", "neutral", "contradiction". We merge the "neutral" and "contradiction" class into one "not-entailment" class to obtain the universal binary format. As figure 5.1 shows, only the probabilities for the "entailment" class are relevant for universal classification. We merge all five NLI datasets into one harmonized dataset with three columns: "premise", "hypothesis", "label".

The resulting merged ~885000 hypothesis-premise pairs would be enough to train a decent NLI model capable of zeroshot classification. The NLI datasets were, however, not created with zeroshot classification in mind. Crowd workers were instructed to write hypotheses that are entailed, contradictory or neutral towards a text, which led to a wide range of hypothesis-premise pairs. They were not specifically instructed to create data for typical classification tasks such as identifying topics, sentiment, stances, emotions, toxicity, factuality etc. which users might be interested in in practice (e.g. "This text is about topic X"). To improve performance on these types of tasks, we therefore add a second collection of standard non-NLI classification datasets reformatted into the NLI format.

[4] `https://github.com/MoritzLaurer/zeroshot-classifier`

[5] We exclude the large SNLI dataset Bowman et al. (2015) due to known issues of data quality.

`data-harmonization-huggingface.ipynb`

We choose 28 popular non-NLI datasets with diverse classification tasks linked to sentiment, emotions, intent, toxicity, bias, topics, factuality, spam etc. with 387 classes in total. We selected most datasets based on their popularity (downloads) on the Hugging Face Hub. We also add some non-NLI datasets that are not available on the Hugging Face hub and create separate preprocessing notebooks for each of them (e.g. `1-data-harmonization-manifesto.ipynb`). The full list of datasets with information on tasks, licenses and data quality is available in our dataset overview file.[6]

For creating this kind of collection, we strongly recommend manually inspecting each dataset and the corresponding paper to understand data quality and the underlying task. Depending on the datasets, the preprocessing steps can include: removing NAs, deduplication, downsampling majority classes, merging texts (e.g. titles with text bodies), converting continuous labels into simpler classes (e.g. star ratings to binary sentiment classes), removing texts with low certainty or annotator agreement, splitting datasets with multiple implicit tasks into separate tasks, removing and renaming columns, and splitting the data into a 80-20 train-test split if no test-set exists. As a result of these steps, each processed dataset only has three harmonized column: "text", "label_text" (a word expressing the meaning of each class), and "label_standard" (a number for each class).

If readers want to improve the classifier on a specific domain or a family of other tasks, they can add their datasets during this step.

5.3.2 Automatic data cleaning

`data-cleaning.ipynb`

Manual inspection of the non-NLI datasets reveal relevant quality issues in many datasets. We therefore use the `CleanLab` library to remove texts with a high probability of noise.[7] `CleanLab` provides automated means for identifying noisy labels by embedding texts with a SentenceBERT model, training a simple logistic classifier on these embeddings and analysing prediction uncertainty and prediction overlaps between classes.

[6] `https://github.com/MoritzLaurer/zeroshot-classifier/blob/main/v1_human_data/datasets_overview.csv`

[7] `https://github.com/cleanlab/cleanlab`

Two relevant limitations of this process are that it can disproportionately remove minority classes and it probably does not work well for very complex tasks. We therefore applied this automatic approach to 25 tasks, but not to complex tasks like NLI or factuality detection. This process removes roughly 17% (or ~135 000) texts with probable misclassifications or label overlaps. We highly recommend readers to inspect our cleaning notebook to get a feeling for the amount of noise that is still present in established datasets.

As an additional measure to increase data quality and diversity in the following script, we also radically downsample data for each non-NLI dataset. We only take a sample of maximum 500 texts per class and maximum 5000 texts per dataset to avoid overfitting to a specific large dataset. This leads to 51731 non-NLI texts (down from more than one million texts) that will be merged with the ~885000 NLI texts in the following step. We could have added hundreds of thousands of additional texts, but our experience indicates that data diversity and quality is more important than quantity. Moreover, our objective is not to build a classifier that beats (and overfits to) a benchmark, but to build a classifier that generalizes well.

5.3.3 Hypothesis formulation and NLI formatting

`data-formatting-universal-nli.ipynb`

We now need to transform the (cleaned) non-NLI datasets into the universal NLI format. First, we need to verbalise each class as a class hypothesis. For this label verbalisation step we read the underlying paper or annotator instructions for each dataset and express them as a class hypothesis. For a binary sentiment classification task on app reviews, for example, the hypotheses could be "This app review text expresses positive sentiment" and "This app review text expresses negative sentiment". We add information on the domain or type of dataset ("app review text") in some hypotheses, to help the model differentiate between texts from the same task type (e.g. binary sentiment classification) that come from different domains or datasets (e.g. app reviews vs. movie reviews vs. product reviews). This helps reduce negative transfer risks across datasets. As a general rule, we try to formulate the hypotheses in simple every-day language and avoid complex academic definitions, thinking of the model a bit like a simple crowd worker. Each class hypothesis is linked to its corresponding class label in a dictionary. All our hypotheses

are available in `3-data-formatting-universal-nli.ipynb`.[8]

For each row in each non-NLI *training* dataset we now add a new "hypothesis" column with the correct class hypotheses corresponding to the respective text. Moreover, in a new "label" column, these text-hypothesis pairs receive the label "0" for "entailment". We then multiply each text by two and pair the copied text with a random incorrect class hypothesis and the label "1" for "not-entailment". This multiplication ensures that the model does not only learn that class hypotheses are always true and it functions as a form of data augmentation. When we rename the "text" column to "premise", this dataset now has exactly the same format as the NLI dataset with the columns "premise", "hypothesis", "label" for binary entailment vs. not-entailment classification. This conversion is implemented in the function `format_nli_trainset`. We can now simply concatenate the non-NLI and the NLI training data.

The non-NLI *test* data needs to be formatted slightly differently. During test-time, all class hypotheses for a task need to be tested on each text to select the "most entailed" hypothesis. This means that we need to multiply each test text by N for N classes, pairing the text with all N possible class hypotheses in N rows. This conversion is implemented in the function `format_nli_testset`. After this task-specific multiplication, these test sets cannot be concatenated and they need to be evaluated separately.

5.3.4 Training and evaluation

`train-eval.ipynb`

With the data fully cleaned and formatted, we can now start training. We can use any pre-trained transformer model as the foundation. Since the only purpose of the model is classification, we discard models with a decoder such as T5 or Llama-2 (Raffel et al., 2020; Touvron et al., 2023). Among encoder-only models, we had the best experience with DeBERTaV3 which is pre-trained with the highly effective RTD objective and exists in multiple sizes and with a multilingual variant (P. He et al., 2021). Processing and training is implemented with Hugging Face `Transformers`. We use `label2id = {"entailment": 0, "not_entailment": 1}` for compatibility with the `ZeroShotClassificationPipeline`; pad and truncate to a maxi-

[8] Research indicates that providing multiple different instructions (hypotheses) for the same class can help increase generalisation (Sanh et al., 2022).

mum length of 512 tokens; base hyperparameters on the recommended fine-tuning hyperparameters in the appendix of the DeBERTaV3 paper (P. He et al., 2021) and do not conduct a hyperparameter search as it adds little value over the recommended hyperparameters in our experience while adding complexity.

We fine-tune models with three different data compositions for evaluation: (1) one model trained on all datasets (`deberta-v3-zeroshot-v1.1-all-33`); (2) one model trained on only the five NLI datasets as a baseline representing previous NLI-only zeroshot models (`deberta-v3-nli-only`); (3) 28 different models, each trained with all datasets, except one non-NLI dataset is held out. This last group of models is trained to test zeroshot generalisation to tasks the model has not seen during training. For each of the 28 models, we take the performance metric for the dataset that was held out in the respective training run. Based on these 28 metrics, we know what the performance for each task would be, if the model had seen all datasets, except the respective held out dataset.

One training run on around 9000000 concatenated hypothesis-premise pairs for 3 epochs takes around 5 hours for DeBERTaV3-base and 10 hours for DeBERTaV3-large on one A100 40GB GPU. Training and evaluating all 30 models takes around 6 (base) or 15 (large) full days of compute, mostly due to the 28 models trained for held-out testing.

We use balanced accuracy as our main evaluation metric (Buitinck et al., 2013) as many of our datasets are class imbalanced and the metric is easier to interpret than F1 macro. For evaluation on non-NLI datasets, remember that rows have been multiplied with one row per class hypothesis. The `compute_metrics_nli_binary` function handles the calculation of metrics for these reformatted datasets.

`deberta-v3-zeroshot-v1.1-all-33` is the model we recommend for downstream use. The model is available in different sizes in our zeroshot collection on the Hugging Face Hub.[9]

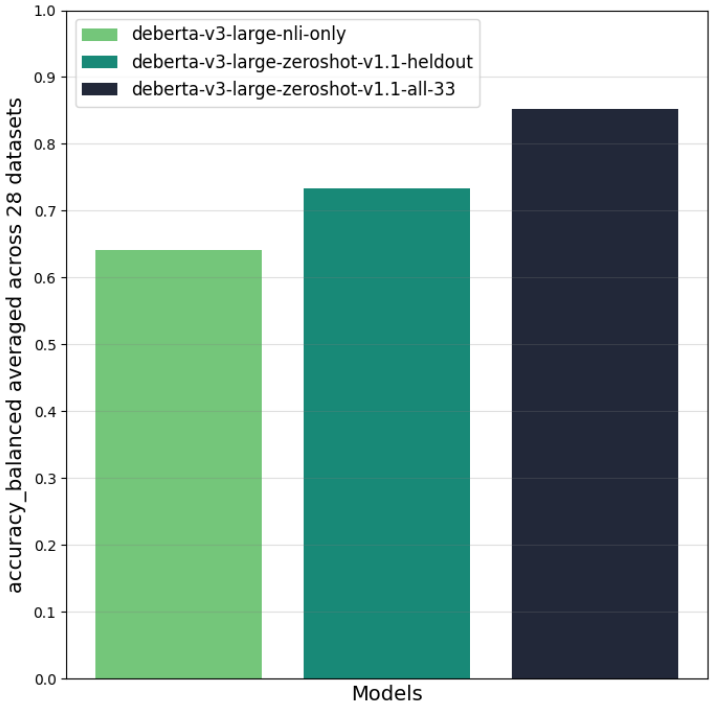
5.3.5 Visualisation and interpretation of results

`viz.ipynb`

The NLI-only classifier (`deberta-v3-nli-only`) is very similar to existing zeroshot classifiers on the Hugging Face hub. It can do all tasks to

[9] `https://huggingface.co/collections/MoritzLaurer/zeroshot-classifiers-6548b4ff407bb19ff5c3ad6f`

Figure 5.3: Mean performance across 28 classification tasks.



some extent, given it's training on the universal NLI task. It performs well on simple binary tasks such as sentiment classification, but struggles on other tasks that are too dissimilar from standard NLI texts and have more classes.

`deberta-v3-zeroshot-v1.1-all-33` has seen up to 500 examples for each class in each dataset. Only based on this small amount of data, it achieves strongly improved performance across all tasks. This is in line with prior research indicating that little, but good quality data is necessary for language models to generalize well (Zhou et al., 2023).

`deberta-v3-zeroshot-v1.1-heldout` provides an indication of zeroshot performance for tasks the model has not seen during training. We highlight two main insights: First, models trained with a mix of NLI data and non-NLI data achieve overall better zeroshot performance than the NLI-only model (+9.4% on average). Having seen different zeroshot-style hypotheses helps the model generalize to other unseen tasks and hypotheses (positive transfer). Second, there are a few cases

of negative transfer. On a few datasets, the NLI-only model performs better than `deberta-v3-zeroshot-v1.1-heldout`, indicating that the additional task-mix can make the model over- or underpredict a few classes.

Overall, `deberta-v3-zeroshot-v1.1-all-33` significantly outperforms the NLI-only model both on held-in and held-out tasks. Its performance on datasets it has not seen during training can expected to be around 9.4% higher than NLI-only models. Moreover, it can simultaneously perform many different tasks it has seen during training with even better performance. Detailed metrics are available in the appendix and the model cards.[10]

5.4 Reusing our models and code

We envisage three main ways in which our models and code can be reused. First, users can directly use `deberta-v3-zeroshot-v1.1-all-33` for zeroshot classification in just a few lines of code with the Hugging Face ZeroShotClassificationPipeline (see code in figure 5.2). This should work particularly well for tasks that are similar to one of the 33 datasets and 389 classes we used for training, including many different topics, sentiment, emotions, or types of toxicity.

Second, the models can be used as a base models to fine-tune a task-specific classifier. Prior research shows that fine-tuning an NLI-based classifier requires less training data and increases robustness compared standard fine-tuning of DeBERTaV3-base (Laurer et al., 2023a; Le Scao & Rush, 2021; Raman et al., 2023). Good performance can be achieved with just a few hundred examples per class, requiring only some minutes of fine-tuning on a free GPU (Laurer et al., 2023b). We provide code examples for this approach in an online workshop.[11].

Third, researchers can modify our notebooks, for example by adding more datasets for a specific domain and task family, and rerun the improved pipeline to build a universal classifier that is better adapted to their domain and tasks. While fine-tuning `deberta-v3-zeroshot-v1.1-all-33` is recommended for individual tasks, rerunning the pipeline could add value if researchers want to build a new universal model adapted to

[10] `https://huggingface.co/MoritzLaurer/deberta-v3-large-zeroshot-v1.1-all-33`

[11] See the notebook `4_tune_bert_nli.ipynb` at `https://github.com/MoritzLaurer/summer-school-transformers-2023/tree/main`

a broader set of tasks or domains. We estimate that the final model can be trained with a € 50 Google Colab Pro+ subscription.

In all three use-cases, making predictions with the resulting models (inference) is highly efficient with cheap GPUs, but is also possible with on a laptop CPU.

5.5 Limitations

We outline several limitations of this paper and invite readers to improve on our implementation. First, while we have included 28 non-NLI datasets, the diversity of these academic datasets is limited and they do not cover the full diversity of classification use-cases users will need in practice. All datasets are only in English. The instruction fine-tuning literature for generative LLMs has shown the potential of using SotA models like GPT-4 to generate diverse training data and distilling their capabilities into much smaller models (Taori et al., 2023; Tunstall et al., 2023). While many such datasets exist for generative tasks, hardly any are available for encoder-only classifiers like BERT (Longpre, Hou, et al., 2023; Longpre, Mahari, et al., 2023; Sileo, 2023). We assume that smart LLM prompting could result in a more diverse dataset than our collection and could further improve generalisation.

Second, the model comparisons are limited as we only compare BERT-NLI models among each other. We do not compare classification performance, inference speed, memory requirements, and costs to larger generative LLMs or APIs.

Third, we assume that our data still contains a certain degree of noise. Additional data cleaning techniques could be used, for example discarding training data where the DeBERTa-v3 model still disagrees with the label after fine-tuning or targeted manual inspection enabled by active learning.

Fourth, an inherent limitation of NLI for zeroshot classification is that each additional class requires an additional forward pass (prediction) through the model. This makes the approach less suitable for tasks with a high amount of classes. At the same time, even if multiple forward passes are required, encoder-only models with only around a hundred million parameters are still more efficient than decoder models with multiple billion parameters while possibly being more accurate Schick and Schütze (2021b); H. Xu et al. (2023).

Fifth, we use the relatively old DeBERTa-v3 from November 2021

(P. He et al., 2021), which misses relevant recent innovations like new positional embeddings like RoPe or AliBi to enable longer context windows (Press, Smith, & Lewis, 2022; Su et al., 2023). Unfortunately we are not aware of a better encoder-only model and releases have recently been dominated by larger generative decoder models.

Sixth, several other universal classification approaches exist that were beyond the scope of this paper: PET, which combines masked-language-modeling and label verbalisation (Schick & Schütze, 2021a), replaced-token-detection combined with prompts (Xia et al., 2022; H. Xu et al., 2023; Y. Yao et al., 2022), question-answering (Bragg et al., 2021), or next-sentence-prediction as an interesting self-supervised alternative to NLI (Ma et al., 2021; Sun et al., 2022).

5.6 Conclusion and call for a new foundation model

This paper explains how to use the Natural Language Inference task to build a universal classifier and provides practical guidance to users. Looking forward, we believe that there is significant room for improvement by building upon the insights from generative LLM research for more efficient classifiers.

First, generative LLMs gain their power by learning their universal task (next-token-prediction) already during self-supervised pre-training and not only during fine-tuning (a limitation of our models). It is possible that universal self-supervised tasks exist for classification tasks as well (or discriminative tasks more generally). The most promising candidate is ELECTRA's replaced-token-detection (RTD) objective (Clark, Luong, Le, & Manning, 2020), which can make models with only a few hundred million parameters perform comparably to models with 1.5 billion parameters that are trained on the the less efficient generative masked-language-modeling objective (P. He et al., 2021). We hypothesize that the RTD objective could be supplemented with a binary "original text" vs. "not-original text" objective, resulting in a universal classification head similar to the universal "entailment" vs. "not-entailment" task - without requiring supervision. H. Xu et al. (2023) go in this direction, but did not experiment with a self-supervised task.

Second, a new foundation model trained on this task could then also be trained with other more recent innovations, which existing encoder-only models are currently lacking: grouped-query attention (Ainslie

et al., 2023), flash attention (Dao, Fu, Ermon, Rudra, & Ré, 2022), better positional embeddings like RoPe or AliBi to enable longer context windows (Press et al., 2022; Su et al., 2023), and scaling pre-training data and compute while only moderately scaling model size for inference-time efficiency (Hoffmann et al., 2022).

Third, similar to generative LLMs, better instruction data could make universal classifiers more useful. As discussed in the limitations section, especially synthetic data from much larger generative LLMs tailored to universal classifiers has the potential to flexibly teach efficient classifiers more diverse and more practically relevant tasks. The creators of the WANLI dataset have already demonstrated this potential with GPT3 (Liu et al., 2022) and it is safe to assume that newer generators will produce even better data.

These points would entail pre-training a new foundation model from scratch, which requires large amounts of resources. We believe that such a foundation model for text classification would be a useful addition to the open-source ecosystem as the field has progress significantly since the last encoder-only models were released and classification tasks constitute a relevant share of both academic and practical applications for language models.

Appendix

The appendix is available in the online preprint: *Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2023). Building Efficient Universal Classifiers with Natural Language Inference (arXiv:2312.17543). arXiv. https://doi.org/10.48550/arXiv.2312.17543*

Additional metrics and details are also reported in the corresponding model cards.[12]

For details on all datasets used, see the overview table.[13] To give citation credit to the authors of all datasets, here is the full list of dataset sources: Adams et al. (2017); Almeida, Hidalgo, and Yamakami (2011); Burst et al. (2020); Casanueva, Temčinas, Gerz, Henderson, and Vulić (2020); Chatterjee, Narahari, Joshi, and Agrawal (2019); Davidson, Warmsley, Macy, and Weber (2017); Faruqui and Das (2018); FitzGerald et al. (2023); Gekhman, Herzig, Aharoni, Elkind, and Szpektor (2023); Grano et al. (2017); Liu et al. (2022); Maas et al. (2011); Malo, Sinha, Korhonen, Wallenius, and Takala (2014); Mathew et al. (2021); McAuley and Leskovec (2013); Nie et al. (2020); Pang and Lee (2005); Parrish et al. (2021); Project (2015); Rashkin, Smith, Li, and Boureau (2019); Sap et al. (2020); Saravia, Liu, Huang, Wu, and Chen (2018); Soups (2015); Thorne, Vlachos, Christodoulopoulos, and Mittal (2018); Unknown (2024); Williams et al. (2018); Zhang, Zhao, and LeCun (2015)

[12] `https://huggingface.co/MoritzLaurer/deberta-v3-large-zeroshot-v1.1-all-33`

[13] `https://github.com/MoritzLaurer/zeroshot-classifier/blob/a7934b8c6c9a5a37764bb4e893dc34bd57587bc2/v1_human_data/datasets_overview.csv`

Chapter 6

Conclusion

In this thesis, I investigated how recent innovations from the Natural Language Processing literature can help address common textual measurement problems in the Computational Social Sciences. More specifically, I showed how the instruction-based language model BERT-NLI leads to better measurements created with text classification.

The uptake of texts classification and supervised machine learning as a measurement tool was limited by several shortcomings of established algorithms: (1) They require large amounts of balanced training data to work well. Researchers, however, often only have limited resources for creating training data and need to tailor data and models to their specific research interest. (2) Older algorithms struggle with multilingual data. Researchers, however, often need measurements that are equally valid for different languages and cultures. (3) They are susceptible to learning shortcuts and biased patterns from their training data, reducing the validity of measurements across social groups. (4) Both older and newer models can be difficult to use in practice, making them only accessible to specialized researchers.

I show that combining language models with instructions provides a solution to these problems. Instruction-based models reuse both prior language and task knowledge (Lou et al., 2023; Ruder, 2019) by building upon universal tasks from pre-training or pre-fine-tuning. They do not need to delete and relearn task-specific parameters. Moreover, their universal tasks enable them to ingest instructions as a third input besides training texts and labels. These verbalised descriptions of tasks, also called prompts, enable them to learn new tasks faster and more robustly.

I empirically tested these assumptions on one type of instruction-

based model: BERT-NLI (Yin et al., 2020). While being limited to text classification, its specific design for text classification makes it an efficient and accessible choice compared to larger generative models (Schick & Schütze, 2021a; H. Xu et al., 2023). I developed my arguments on the benefits of the instruction-based BERT-NLI model in four chapters.

Chapter 2 demonstrated the benefits of BERT-NLI on a wide range of eight social science classification tasks. Across these eight tasks, BERT-NLI fine-tuned on 100 to 2500 texts performs on average 10.7% to 18.3% (percentage points) better than classical models that do not use transfer learning. Results indicate that BERT-NLI fine-tuned on 500 texts achieves similar performance as classical models trained on around 5000 texts. Moreover, I show that BERT-NLI also performs well on imbalanced data.

Chapter 3 investigated the extent to which prior language and task knowledge stored in the parameters of modern language models is useful for enabling multilingual research. Moreover, I test to what extent these algorithms can be fruitfully combined with machine translation. The experiments are designed to determine which methods are most accurate, practical, and valid in multilingual settings – three essential conditions for lowering the language barrier in practice. The analysis is conducted on two datasets with texts in 12 languages from 27 countries and shows that instruction-based models can produce substantively meaningful outputs. Our BERT-NLI model trained on only 674 or 1,674 texts in only one or two languages can validly predict political party families' stances towards immigration in eight other languages and ten other countries.

Chapter 4 investigated the effects of group-specific biases in the training data. This risk of hidden biases is 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 groups. To investigate this issue, I conducted a comparative analysis across nine group types in four datasets with three types of classification models, focusing on the robustness of models against group-specific biases and the validity of their outputs. I find that all types of models learn group-specific biases. On average, however, the 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 produce more valid measurements across different groups. The instruction-based BERT-NLI

sees its average test-set performance decrease by only 0.4% F1 macro when trained on biased data and its error probability on groups it has not seen during training increases only by 0.8%.

Chapter 5 demonstrated how BERT-NLI can be used as a universal classifier. While generative LLMs are an increasingly popular universal text analysis tool, I argue that users do not necessarily need the capability overhang of a generative LLM when they only want to automate a classification task. This chapter demonstrates that Natural Language Inference (NLI) is a viable alternative universal task that can be used to address any text classification task in a zeroshot manner with models significantly smaller than generative LLMs. Each section of the chapter is accompanied by a Jupyter notebook that implements each step required to train such a universal classifier. The resulting model is trained on 33 datasets with 389 classes simultaneously, covering both NLI data and non-NLI data reformatted into the universal NLI format. Its performance at doing new classification tasks without having seen training data (zeroshot classification) increases by 9.4% balanced accuracy compared to NLI-only models. It can run for inference on a laptop and can be reproduced in the browser for around 50 Euros on Google Colab.

In summary, my thesis demonstrated the added value of instruction-based language models using the example of BERT-NLI. BERT-NLI can significantly reduce the required amount of training data and handles imbalanced data well; it can produce valid measurements in multilingual settings; it is more robust against biases from group-specific language patterns than classifiers without instructions; and I provide a range of Jupyter notebooks and fine-tuned models to facilitate the use of these models for settings with little or no training data. On the day of submission of this thesis, my open-source BERT-NLI models have been downloaded more than 65 million times.[1]

6.1 Discussion

My thesis has, however, several limitations and there are many research directions that can further address different practical challenges.

First, my thesis only focused on one type of instruction-based model: BERT-NLI. This was a conscious choice given that it is specialised in text classification, the key task family for social science measurement.

[1] 🤗 `https://huggingface.co/MoritzLaurer`

Generative instruction-based models also have great potential for simplifying measurement beyond classification, but they were not covered in this thesis given their accessibility issues in terms of size and costs for academics during the time of writing. There are also other universal tasks for instruction-based classification models that could have been explored (see chapter 5, section 6).

Moreover, I did not investigate procedures for better training data sampling and annotation. I mostly sampled randomly from existing annotated datasets, but practitioners are normally faced with a raw corpus in which specific unbalanced classes need to be identified. I believe that two procedures have great potential for simplifying this process. First, active learning can be an efficient means for sampling data (Schröder & Niekler, 2020), especially in imbalanced settings (Miller et al., 2020). Classifiers like BERT-NLI can be fruitfully combined with active learning, as they provide decently calibrated uncertainty scores with little or no training data. I have started combining BERT-NLI with active learning for an open-source tutorial [2] , but more rigorous experiments would be necessary to establish advantages and disadvantages in different settings. Second, the creation of synthetic data with large generative models is promising. Recent evidence suggests that large language models like GPT-4 can create data that is of similar or better quality as data created by crowd workers (Gilardi, Alizadeh, & Kubli, 2023; X. He et al., 2023; Zheng et al., 2023). State-of-the-Art LLMs can be used to annotate texts with good zeroshot performance or they can be instructed to write new texts tailored to a specific research question. This data can be used as training data for much smaller models ("knowledge distillation", Sanh, Debut, Chaumond, & Wolf, 2020; Taori et al., 2023). I demonstrate the value of synthetic data in a blog post for practitioners outside of my thesis.[3]

Lastly, there are many risks associated with languages models which are not discussed in my thesis. Only chapter 4 investigated the issue of biases in training data, but many more risks exist. Besides measurement tools, language models can also be tools for discrimination, mass surveillance, disinformation, manipulation, or could become misaligned autonomous agents (Bender et al., 2021; OpenAI, 2023a; Weidinger et al., 2021). Moreover, the training and use of LLMs are very energy intensive.

[2] https://github.com/argilla-io/argilla/blob/61967367606724dfa8e0b25d1ab2185232d59b73/docs/_source/tutorials/notebooks/deploying-textclassification-colab-activelearning.ipynb

[3] https://huggingface.co/blog/synthetic-data-save-costs

In the blog post mentioned above, for example, I demonstrate that an analysis with GPT-4 can emit around 735 to 1100 kg CO2 equivalents, while an analysis with a smaller BERT classifier only emits around 0.12 kg CO2. Interdisciplinary researchers should continue researching the full spectrum of risks. Social scientists have a particularly important role to play in researching hidden biases or misuse and propose practical solutions both on a technical and a governance level (Irving & Askell, 2019).

6.2 Looking ahead

Given the speed at which language models and related technologies have evolved in the past few years, it is important to look ahead and assess how the landscape might change in the near to medium term. I started my thesis in 2021, when generative LLMs were still mostly out of reach for academics. Since 2023, anyone can use an LLM with a simple API call, achieving impressive performance without having to think about fine-tuning or hardware. How could measurement with language models evolve in the coming three years?

First, deriving measurements from text will become dramatically easier and new instruction-based models will play an essential role in this development. The most prominent precursor for this is the GPT-4 API (OpenAI, 2023b). GPT-4 is probably already better than crowd workers at most text annotation tasks, while being cheaper and easier to interact with (Gilardi et al., 2023; X. He et al., 2023; Zheng et al., 2023). I anticipate that instruction-based language models will soon be like on-demand research assistants whose (artificial) intelligence can be bought much more flexibly and cheaply than the intellectual labour of human workers at similar or better performance. The complex task of fine-tuning language models will become less important for many users and their main interaction with models will be the iterative testing of instructions and validation of outputs. Simple API calls or user interfaces will significantly reduce the expertise required to use these artificial research assistants. While the fine-tuning process might take weeks today, the iterative prompting process might only take days. This will make the creation of measurements from large text corpora dramatically easier in the near to medium term. Future research will need to develop best practices for using LLMs to avoid that their ease-of-use leads to an erosion of good existing validation practices.

Second, among instruction-based model types, especially generative variants will open new possibilities to create measurements beyond the classical categories of classification or scaling (Grimmer & Stewart, 2013). New methods for measurement could include: (a) exploratory categorisation that is less restricted to a pre-defined set of classes than supervised classification, but more directed than unsupervised clustering. By writing instructions that specify a specific research interest, but do not restrict the generative model to a set of categories, researchers will be able to explore a new middle-ground between inductive and deductive research. (b) New forms of scaling and numeric measurements will emerge. While classification puts text into discrete categories, scaling assigns texts a continuous numeric score, such as a positioning on an ideological scale. Generative models are increasingly used to rate texts with continuous numeric variables (Zheng et al., 2023). Social scientists will be able to use this capability to create diverse continuous ratings of texts and guide generative models to produce numeric ratings like degrees of emotions, beliefs, or text quality (Wu, Nagler, Tucker, & Messing, 2023). Instructions could be used, for example, to implement new forms of ideological scaling that is more strongly based on the researcher's interest and less dependent on potentially spurious patterns in the data. (c) Generative models can be used for entirely new experimental designs, such as using LLMs to simulate human samples or create controlled experiments where people interact with a carefully prompted LLM instead of researchers (Argyle et al., 2023).

Moreover, (d) generative models will help combine previously separated analysis steps into one unified framework. A key limitation of most computational text analysis methods today is that a different computational tool is required for each subtask in a larger analysis ("atomization", Baden et al., 2022). An analysis might, for example, require identifying any politician mentioned in a text, extracting their name strings, harmonizing these strings (e.g. "Dr. Merkel" and "the leader of the CDU in 2014" to "Angela Merkel"), and determining the stance towards each specific politician mentioned in the text. Established text-as-data methods would require a complex chain of different methods from named-entity-extraction, entity-linking, to separate stance classifiers. Generative models can, in theory, solve this task in a unified conversational framework. A first instruction can ask the model to extract all mentions of politicians in a structured list. A second instruction can ask it to harmonise the list based on an existing list of

names. A third instruction can ask it to output the stance towards each person in structured JSON format. These instructions can build-upon each other in a conversation, where later responses are conditioned on previous responses. Combining the chat design of generative models with "chains-of-thought" has great potential for text analyses (Wei et al., 2023). While there are important challenges (cascading errors and the need to validate each subtask), the universality of the text generation task and the chat design of LLMs have great potential to overcome the "Specialization before Integration" gap in text-as-data research (Baden et al., 2022).

Third, as more analyses become possible, the value of raw data that was previously too difficult to analyse will increase. The most important frontier is probably multimodal research, where different modalities like text, image, video, or audio are ingested simultaneously by a multimodal language model. A prime use-case is social media data, where the full meaning of a post can only be understood by combining text, images, and audio. Over the past years, the machine learning literature has been busy developing methods that can ingest data in multiple modalities simultaneously (P. Xu, Zhu, & Clifton, 2023). Several open-source models are available on the Hugging Face platform that can be directly fine-tuned (Wolf et al., 2020) and APIs provide easy ways to use generalist multimodal models (OpenAI 2023). Large scale data collection projects that collect (multimodal) data with these new capabilities in mind and multimodal research in general will become more important.

Fourth, as these methods and data become available, tooling and methods education in the social sciences need to adapt (Van Atteveldt & Peng, 2018). The most important stumbling block might be the lack of programming education in the social sciences, especially in the Python programming language. The R community in the social sciences has built an amazing ecosystem for statistical data analyses, text-as-data research, visualisation and more. In parallel, deep learning researchers have built almost all recent innovations in Python and practitioners have built an amazing ecosystem of libraries that make these innovations easy to use. Uptake of these innovations is slowed down significantly, not because the innovations are more difficult to use, but because they are implemented in Python and computational social scientists tend to be more comfortable in R.

There are several pathways that can make tools from the deep

learning community more accessible for social scientists. (a) Teaching of programming languages in the computational social sciences should become more prominent and more diverse. Python should become an equal alternative to R in social science textbooks and curricula to enable students to make an informed decision about which language is more suitable for their research. (b) New tools like GitHub Copilot or ChatGPT will make it significantly easier to learn a new programming language or tool. While in 2022, solving a problem in a new language required skimming long documentation and finding suitable responses in online fora, learners can now simply copy an error message into a chat window and receive good quality responses tailored to their problem. Moreover, code from one programming language can be translated into another language with decent quality and these services will only become better. (c) APIs to use generalist models will make the choice of programming language less relevant. APIs can be called in any language and the outputs can then be processed in the language the researcher is comfortable with. Overall, it has never been easier to learn programming and to use advanced models. As programming becomes both an increasingly relevant tool and an increasingly relevant subject of study, social science education should invest more in empowering students to understand and use these tools.

Fifth, compute infrastructure will (have to) become more accessible. For a relevant share of research projects, APIs will not be enough. Researchers can now access advanced hardware for free or a small fee in their browser via services like Google Colab and these services will improve. For more compute intensive research, universities and public bodies should cooperate and invest in public compute infrastructure. Cooperative projects like the Dutch supercomputer Snellius are immensely valuable for enabling research on newer models.

Lastly, despite the excitement about the capabilities of new LLMs, it is essential to remember that creating valid measurements requires much more than building machine annotators with human-level performance. As discussed in chapter 4, four interdependent steps are required for creating a valid measurement (Adcock & Collier, 2001): (1) conceptualizing a general background concept into a specific systematized concept; (2) operationalizing the systematized concept in a meaningful indicator; (3) scoring individual cases based on this operationalization (e.g. text classification); (4) aggregating the individual scores into the indicator, i.e. the final 'measurement'. Today's LLMs can support these steps

individually if certain conditions are met: They can achieve human-level performance for scoring individual texts, if the underlying concept, data and instructions are meaningful; they can be expert assistants for brainstorming on conceptualization and operationalization if prompted properly; and they can assist in writing code for aggregation if the input data is meaningful.[4] Reliably and validly combining these steps to create a good measurement end-to-end, however, is still out of reach for today's best models.[5] Most importantly, even once a perfectly valid measurement is created, it is still only a means to an end: understanding and explaining society. After all, a good measurement will only be one input variable in a broader explanatory model (Egami et al., 2022; Grimmer et al., 2021; Wallach, 2018).

The coming years hold many exciting research opportunities for interdisciplinary research between machine learning and the social sciences. I hope that many researchers will explore how these new tools can help solve different practical problems in research and beyond. As language models establish themselves as a valuable addition to our toolbox, different disciplines will learn more about their opportunities and limitations. I hope that my thesis made a contribution to this endeavour.

[4]Implementing these steps is not only a question of capability or intelligence, but also of research interest and values. Interests and values will always remain contested and navigating them will remain an important challenge for human experts.

[5]For a machine learning system to start automating scientific research, it needs the ability to plan multiple complex steps and act on each step without cascading errors. These types of agentic language model systems are an active area of research that is already receiving dramatic investments with initial promising results (Mialon, Dessì, et al., 2023; Nakano et al., 2022; Qin et al., 2023; Schick et al., 2023; S. Yao et al., 2023). At least today's systems are, however, still incapable of relatively simple multi-step chains of planning and action (Mialon, Fourrier, et al., 2023). For the purpose of measurement, today's best language models are not much more than (accurate and scalable) text analysis machines. It is reasonable to expect, however, that their capabilities will increase dramatically in the short to medium term.

Bibliography

Adams, C., Cukierski, W., Sorensen, J., Elliott, J., Dixon, L., McDonald, M., & nithum. (2017). *Toxic Comment Classification Challenge.* Kaggle. Retrieved from https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge

Adcock, R., & Collier, D. (2001, September). Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review*, *95*(3), 529–546. Retrieved 2023-01-13, from https://www.cambridge.org/core/journals/american-political-science-review/article/measurement-validity-a-shared-standard-for-qualitative-and-quantitative-research/91C7A9800DB26A76EBBABC5889A50C8B (Publisher: Cambridge University Press) doi: 10.1017/S0003055401003100

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., & Sanghai, S. (2023, October). *GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints.* arXiv. Retrieved 2023-11-14, from http://arxiv.org/abs/2305.13245 (arXiv:2305.13245 [cs])

Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011, September). Contributions to the study of SMS spam filtering: new collection and results. In *Proceedings of the 11th ACM symposium on Document engineering* (pp. 259–262). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-03-22, from https://doi.org/10.1145/2034691.2034742 doi: 10.1145/2034691.2034742

Anastasopoulos, L. J., & Bertelli, A. M. (2020, February). Understanding Delegation Through Machine Learning: A Method and Application to the European Union. *American Political Science Review*, *114*(1), 291–301. Retrieved 2022-04-30, from http://www.cambridge.org/core/journals/american-political-science-review/article/

understanding-delegation-through-machine-learning-a-method-and-application-to-the-european-union/1724F3ECFA1F0AABE3C7F8DA5C5D521B (Publisher: Cambridge University Press) doi: 10.1017/S0003055419000522

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023, February). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 1–15. Retrieved 2023-02-24, from https://www.cambridge.org/core/journals/political-analysis/article/out-of-one-many-using-language-models-to-simulate-human-samples/035D7C8A55B237942FB6DBAD7CAA4E49?utm_source=SFMC&utm_medium=email&utm_content=Article&utm_campaign=New%20Cambridge%20Alert%20-%20Articles&WT.mc_id=New%20Cambridge%20Alert%20-%20Articles (Publisher: Cambridge University Press) doi: 10.1017/pan.2023.2

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022, January). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, *16*(1), 1–18. Retrieved 2023-09-15, from https://doi.org/10.1080/19312458.2021.2015574 (Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2021.2015574) doi: 10.1080/19312458.2021.2015574

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021, January). Automated Text Classification of News Articles: A Practical Guide. *Political Analysis*, *29*(1), 19–42. Retrieved 2021-12-23, from http://www.cambridge.org/core/journals/political-analysis/article/automated-text-classification-of-news-articles-a-practical-guide/10462DB284B1CD80C0FAE796AD786BC6 (Publisher: Cambridge University Press) doi: 10.1017/pan.2020.8

Barriere, V., & Balahur, A. (2020, October). *Improving Sentiment Analysis over non-English Tweets using Multilingual Transformers and Automatic Translation for Data-Augmentation.* arXiv. Retrieved 2022-11-24, from http://arxiv.org/abs/2010.03486 (arXiv:2010.03486 [cs]) doi: 10.48550/arXiv.2010.03486

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Confer-*

ence on Fairness, Accountability, and Transparency (pp. 610–623). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-05-01, from https://dl.acm.org/doi/10.1145/3442188.3445922 doi: 10.1145/3442188.3445922

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A Neural Probabilistic Language Model. In *Advances in Neural Information Processing Systems* (Vol. 13). MIT Press. Retrieved 2023-11-20, from https://papers.nips.cc/paper_files/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html

Benoit, K. (2020). Text as Data: An Overview. In L. Curini & R. Franzese (Eds.), *The SAGE Handbook of Research Methods in Political Science and International Relations* (pp. 461–497). 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd. Retrieved 2021-12-19, from https://sk.sagepub.com/reference/the-sage-handbook-of-research-methods-in-political-science-and-ir/i4365.xml doi: 10.4135/9781526486387.n29

Bestvater, S. E., & Monroe, B. L. (2022, April). Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis. *Political Analysis*, 1–22. Retrieved 2022-04-29, from http://www.cambridge.org/core/journals/political-analysis/article/sentiment-is-not-stance-targetaware-opinion-classification-for-political-text-analysis/743A9DD62DF3F2F448E199BDD1C37C8D (Publisher: Cambridge University Press) doi: 10.1017/pan.2022.10

Bilbao-Jayo, A., & Almeida, A. (2018, November). Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data. *International Journal of Distributed Sensor Networks*, *14*(11), 155014771881182. Retrieved 2022-03-31, from http://journals.sagepub.com/doi/10.1177/1550147718811827 doi: 10.1177/1550147718811827

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020, May). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv:2005.14050 [cs]*. Retrieved 2022-02-10, from http://arxiv.org/abs/2005.14050 (arXiv: 2005.14050)

Bornea, M., Pan, L., Rosenthal, S., Florian, R., & Sil, A. (2021, May). Multilingual Transfer Learning for QA using Translation as Data Augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(14), 12583–12591. Retrieved 2022-

11-24, from https://ojs.aaai.org/index.php/AAAI/article/view/17491 (Number: 14) doi: 10.1609/aaai.v35i14.17491

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015, August). A large annotated corpus for learning natural language inference. *arXiv:1508.05326 [cs]*. Retrieved 2021-12-24, from http://arxiv.org/abs/1508.05326 (arXiv: 1508.05326)

Bowman, S. R., & Dahl, G. (2021, June). What Will it Take to Fix Benchmarking in Natural Language Understanding? In K. Toutanova et al. (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4843–4855). Online: Association for Computational Linguistics. Retrieved 2023-11-20, from https://aclanthology.org/2021.naacl-main.385 doi: 10.18653/v1/2021.naacl-main.385

Bragg, J., Cohan, A., Lo, K., & Beltagy, I. (2021, November). FLEX: Unifying Evaluation for Few-Shot NLP. *arXiv:2107.07170 [cs]*. Retrieved 2022-01-07, from http://arxiv.org/abs/2107.07170 (arXiv: 2107.07170)

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. Retrieved 2023-05-30, from https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., … Varoquaux, G. (2013, September). *API design for machine learning software: experiences from the scikit-learn project.* arXiv. Retrieved 2023-11-14, from http://arxiv.org/abs/1309.0238 (arXiv:1309.0238 [cs]) doi: 10.48550/arXiv.1309.0238

Burst, T., Werner, K., Lehmann, P., Jirka, L., Mattheiß, T., Merz, N., … Zehnter, L. (2020). *Manifesto Corpus.* WZB Berlin Social Science Center. Retrieved 2021-06-26, from https://manifesto-project.wzb.eu/information/documents/corpus

Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020, July). Efficient Intent Detection with Dual Sentence Encoders. In T.-H. Wen et al. (Eds.), *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved 2024-03-22, from https://aclanthology.org/2020.nlp4convai-1.5

doi: 10.18653/v1/2020.nlp4convai-1.5

Caton, S., & Haas, C. (2020, October). *Fairness in Machine Learning: A Survey.* arXiv. Retrieved 2022-08-03, from http://arxiv.org/abs/2010.04053 (arXiv:2010.04053 [cs, stat]) doi: 10.48550/arXiv.2010.04053

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014, March). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, *16*(2), 340–358. Retrieved 2022-02-05, from https://doi.org/10.1177/1461444813480466 (Publisher: SAGE Publications) doi: 10.1177/1461444813480466

Chatsiou, K., & Mikhaylov, S. J. (2020). Deep Learning for Political Science. *arXiv:2005.06540 [cs]*, *preprint*. Retrieved 2022-01-07, from http://arxiv.org/abs/2005.06540 (arXiv: 2005.06540)

Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019, June). SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 39–48). Minneapolis, Minnesota, USA: Association for Computational Linguistics. Retrieved 2024-03-22, from https://aclanthology.org/S19-2005 doi: 10.18653/v1/S19-2005

Cheng, C., Barceló, J., Hartnett, A. S., Kubinec, R., & Messerschmidt, L. (2020, July). COVID-19 Government Response Event Dataset (CoronaNet v.1.0). *Nature Human Behaviour*, *4*(7), 756–768. Retrieved 2022-03-13, from https://www.nature.com/articles/s41562-020-0909-7 (Number: 7 Publisher: Nature Publishing Group) doi: 10.1038/s41562-020-0909-7

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., … Fiedel, N. (2022, October). *PaLM: Scaling Language Modeling with Pathways.* arXiv. Retrieved 2022-12-29, from http://arxiv.org/abs/2204.02311 (arXiv:2204.02311 [cs])

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., … Wei, J. (2022, December). *Scaling Instruction-Finetuned Language Models.* arXiv. Retrieved 2022-12-21, from http://arxiv.org/abs/2210.11416 (arXiv:2210.11416 [cs])

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020, March). ELECTRA: Pre-training Text Encoders as Discriminators Rather

Than Generators. *arXiv:2003.10555 [cs]*. Retrieved 2022-02-11, from http://arxiv.org/abs/2003.10555 (arXiv: 2003.10555)

Cocco, J. D., & Monechi, B. (2021, October). How Populist are Parties? Measuring Degrees of Populism in Party Manifestos Using Supervised Machine Learning. *Political Analysis*, 1–17. Retrieved 2022-01-09, from https://www.cambridge.org/core/journals/political-analysis/article/how-populist-are-parties-measuring-degrees-of-populism-in-party-manifestos-using-supervised-machine-learning/1D6141AAAE400ADAD9935044A0719B32 (Publisher: Cambridge University Press) doi: 10.1017/pan.2021.29

Colleoni, E., Rozza, A., & Arvidsson, A. (2014, April). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data: Political Homophily on Twitter. *Journal of Communication*, *64*(2), 317–332. Retrieved 2022-02-05, from https://academic.oup.com/joc/article/64/2/317-332/4085994 doi: 10.1111/jcom.12084

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2020, April). Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*. Retrieved 2022-01-05, from http://arxiv.org/abs/1911.02116 (arXiv: 1911.02116)

Conneau, A., & Lample, G. (2019). Cross-lingual Language Model Pretraining. *NeurIPS*, *33*, 11. Retrieved from https://papers.nips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018, September). XNLI: Evaluating Cross-lingual Sentence Representations. *arXiv:1809.05053 [cs]*. Retrieved 2022-01-05, from http://arxiv.org/abs/1809.05053 (arXiv: 1809.05053)

Courtney, M., Breen, M., McMenamin, I., & McNulty, G. (2020, July). Automatic translation, context, and supervised learning in comparative politics. *Journal of Information Technology & Politics*, *17*(3), 208–217. Retrieved 2022-08-02, from https://doi.org/10.1080/19331681.2020.1731245 (Publisher: Routledge _eprint: https://doi.org/10.1080/19331681.2020.1731245) doi: 10.1080/19331681.2020.1731245

Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL Recog-

nising Textual Entailment Challenge. In J. Quiñonero-Candela, I. Dagan, B. Magnini, & F. d'Alché Buc (Eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment* (Vol. 3944, pp. 177–190). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved 2023-11-10, from http://link.springer.com/10.1007/11736790_9 (Series Title: Lecture Notes in Computer Science) doi: 10.1007/11736790_9

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022, June). *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness.* arXiv. Retrieved 2023-06-04, from http://arxiv.org/abs/2205.14135 (arXiv:2205.14135 [cs])

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512–515. Retrieved 2024-03-22, from https://ojs.aaai.org/index.php/ICWSM/article/view/14955 (Number: 1) doi: 10.1609/icwsm.v11i1.14955

Denny, M. J., & Spirling, A. (2018, April). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, *26*(2), 168–189. Retrieved 2023-06-09, from https://www.cambridge.org/core/journals/political-analysis/article/text-preprocessing-for-unsupervised-learning-why-it-matters-when-it-misleads-and-what-to-do-about-it/AA7D4DE0AA6AB208502515AE3EC6989E (Publisher: Cambridge University Press) doi: 10.1017/pan.2017.44

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved 2023-05-30, from http://aclweb.org/anthology/N19-1423 doi: 10.18653/v1/N19-1423

de Vries, E., Schoonvelde, M., & Schumacher, G. (2018, October). No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Political Analysis*, *26*(4), 417–430. Retrieved 2022-07-21, from http://www.cambridge.org/core/journals/political-analysis/article/no-longer

-lost-in-translation-evidence-that-google-translate-works-for-comparative-bagofwords-text-applications/43CB03805973BB8AD567F7AE50E72CA6 (Publisher: Cambridge University Press) doi: 10.1017/pan.2018.26

Du, M., He, F., Zou, N., Tao, D., & Hu, X. (2022, August). *Shortcut Learning of Large Language Models in Natural Language Understanding: A Survey.* arXiv. Retrieved 2023-04-30, from http://arxiv.org/abs/2208.11857 (arXiv:2208.11857 [cs])

Düpont, N., & Rachuj, M. (2022, April). The Ties That Bind: Text Similarities and Conditional Diffusion among Parties. *British Journal of Political Science*, *52*(2), 613–630. Retrieved 2022-07-21, from http://www.cambridge.org/core/journals/british-journal-of-political-science/article/ties-that-bind-text-similarities-and-conditional-diffusion-among-parties/FF43DC2E4A56F2AB5978E2859A5B3F6A (Publisher: Cambridge University Press) doi: 10.1017/S0007123420000617

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022, October). How to make causal inferences using texts. *Science Advances*, *8*(42), eabg2652. Retrieved 2023-03-05, from https://www.science.org/doi/10.1126/sciadv.abg2652 (Publisher: American Association for the Advancement of Science) doi: 10.1126/sciadv.abg2652

Eurobarometer. (2012, June). *Europeans and Their Languages - Report on Special Eurobarometer 386* (Tech. Rep.). TNS Opinion & Social requested by European Commission. Retrieved from https://europa.eu/eurobarometer/surveys/detail/1049

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., … Joulin, A. (2020, October). *Beyond English-Centric Multilingual Machine Translation.* arXiv. Retrieved 2022-11-28, from http://arxiv.org/abs/2010.11125 (arXiv:2010.11125 [cs]) doi: 10.48550/arXiv.2010.11125

Faruqui, M., & Das, D. (2018, October). Identifying Well-formed Natural Language Questions. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 798–803). Brussels, Belgium: Association for Computational Linguistics. Retrieved 2024-03-22, from https://aclanthology.org/D18-1091 doi: 10.18653/v1/D18-1091

FitzGerald, J., Hench, C., Peris, C., Mackie, S., Rottmann, K.,

Sanchez, A., … Natarajan, P. (2023, July). MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4277–4302). Toronto, Canada: Association for Computational Linguistics. Retrieved 2024-03-22, from https://aclanthology.org/2023.acl-long.235 doi: 10.18653/v1/2023.acl-long.235

Gekhman, Z., Herzig, J., Aharoni, R., Elkind, C., & Szpektor, I. (2023, May). *TrueTeacher: Learning Factual Consistency Evaluation with Large Language Models.* arXiv. Retrieved 2023-05-22, from http://arxiv.org/abs/2305.11171 (arXiv:2305.11171 [cs])

Gilardi, F., Alizadeh, M., & Kubli, M. (2023, July). ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. Retrieved 2023-11-19, from http://arxiv.org/abs/2303.15056 (arXiv:2303.15056 [cs]) doi: 10.1073/pnas.2305016120

Grano, G., Di Sorbo, A., Mercaldo, F., Visaggio, C. A., Canfora, G., & Panichella, S. (2017, September). Android apps and user feedback: a dataset for software evolution and quality improvement. In *Proceedings of the 2nd ACM SIGSOFT International Workshop on App Market Analytics* (pp. 8–11). Paderborn Germany: ACM. Retrieved 2024-03-22, from https://dl.acm.org/doi/10.1145/3121264.3121266 doi: 10.1145/3121264.3121266

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021, May). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*(1), 395–419. Retrieved 2023-03-10, from https://www.annualreviews.org/doi/10.1146/annurev-polisci-053119-015921 doi: 10.1146/annurev-polisci-053119-015921

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: a new framework for machine learning and the social sciences.* Princeton Oxford: 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. Retrieved 2021-12-19, from https://www.cambridge.org/core/journals/political-analysis/article/text-as-data-the-promise

-and-pitfalls-of-automatic-content-analysis-methods-for-political-texts/F7AAC8B2909441603FEB25C156448F20 (Publisher: Cambridge University Press) doi: 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. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 107–112). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved 2022-02-10, from http://aclweb.org/anthology/N18-2017 doi: 10.18653/v1/N18-2017

He, P., Gao, J., & Chen, W. (2021, December). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv:2111.09543 [cs]*. Retrieved 2022-01-10, from http://arxiv.org/abs/2111.09543 (arXiv: 2111.09543)

He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., … Chen, W. (2023, March). *AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators.* arXiv. Retrieved 2023-04-08, from http://arxiv.org/abs/2303.16854 (arXiv:2303.16854 [cs])

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). Amsterdam: John Benjamins Publishing Company. Retrieved 2023-02-21, from https://benjamins.com/catalog/dapsac.55.05hir doi: 10.1075/dapsac.55.05hir

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., … Sifre, L. (2022, March). *Training Compute-Optimal Large Language Models.* arXiv. Retrieved 2023-11-14, from http://arxiv.org/abs/2203.15556 (arXiv:2203.15556 [cs]) doi: 10.48550/arXiv.2203.15556

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q. D., Gesmundo, A., … Gelly, S. (2019, May). Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 2790–2799). PMLR. Retrieved 2024-01-05, from https://proceedings.mlr

.press/v97/houlsby19a.html (ISSN: 2640-3498)

Howard, J., & Ruder, S. (2018, July). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 328–339). Melbourne, Australia: Association for Computational Linguistics. Retrieved 2023-05-30, from https://aclanthology.org/P18-1031 doi: 10.18653/v1/P18-1031

Irving, G., & Askell, A. (2019, February). AI Safety Needs Social Scientists. *Distill*, *4*(2), 10.23915/distill.00014. Retrieved 2024-01-13, from https://distill.pub/2019/safety-needs-social-scientists doi: 10.23915/distill.00014

Jacobs, A. Z., & Wallach, H. (2021, March). Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 375–385). New York, NY, USA: Association for Computing Machinery. Retrieved 2022-08-03, from https://doi.org/10.1145/3442188.3445901 doi: 10.1145/3442188.3445901

Jankowski, M., & Huber, R. A. (2023, January). When Correlation Is Not Enough: Validating Populism Scores from Supervised Machine-Learning Models. *Political Analysis*, 1–15. Retrieved 2023-01-10, from https://www.cambridge.org/core/journals/political-analysis/article/when-correlation-is-not-enough-validating-populism-scores-from-supervised-machinelearning-models/8CB3DCBECD1E37872074E7F8A9DE20BA?utm_source=hootsuite&utm_medium=twitter&utm_campaign=PAN_Jan23 (Publisher: Cambridge University Press) doi: 10.1017/pan.2022.32

Kapoor, S., Cantrell, E., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., … Narayanan, A. (2023, September). *REFORMS: Reporting Standards for Machine Learning Based Science.* arXiv. Retrieved 2023-11-20, from http://arxiv.org/abs/2308.07832 (arXiv:2308.07832 [cs, stat]) doi: 10.48550/arXiv.2308.07832

Krippendorff, K. (2018). *Content analysis: an introduction to its methodology* (Fourth Edition ed.). Los Angeles: SAGE.

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023a, June). Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer

Learning and BERT-NLI. *Political Analysis*, 1–33. Retrieved 2023-06-20, from https://www.cambridge.org/core/product/identifier/S1047198723000207/type/journal_article doi: 10.1017/pan.2023.20

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023b, January). Lowering the Language Barrier: Investigating Deep Transfer Learning and Machine Translation for Multilingual Analyses of Political Texts. *Computational Communication Research*, *5*(2), 1. Retrieved 2023-10-31, from https://www.aup-online.com/content/journals/10.5117/CCR2023.2.7.LAUR doi: 10.5117/CCR2023.2.7.LAUR

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023c). *On Measurement Validity and Language Models: Increasing Validity and Robustness against Bias with Instructions.* Retrieved from https://osf.io/hxb5m

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014, March). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. Retrieved 2023-05-01, from https://www.science.org/doi/10.1126/science.1248506 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.1248506

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, *521*(7553), 436–444. Retrieved 2023-11-20, from https://www.nature.com/articles/nature14539 doi: 10.1038/nature14539

Le Scao, T., & Rush, A. (2021). How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2627–2636). Online: Association for Computational Linguistics. Retrieved 2022-03-05, from https://aclanthology.org/2021.naacl-main.208 doi: 10.18653/v1/2021.naacl-main.208

Li, B., Hou, Y., & Che, W. (2022, January). Data augmentation approaches in natural language processing: A survey. *AI Open*, *3*, 71–90. Retrieved 2022-11-24, from https://www.sciencedirect.com/science/article/pii/S2666651022000080 doi: 10.1016/j.aiopen.2022.03.001

Liao, T., Taori, R., Raji, D., & Schmidt, L. (2021, December). Are We Learning Yet? A Meta Review of Evaluation Fail-

ures Across Machine Learning. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, *1*. Retrieved 2023-06-01, from https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/757b505cfd34c64c85ca5b5690ee5293-Abstract-round2.html

Licht, H. (2023, January). Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings. *Political Analysis*, 1–14. Retrieved 2023-04-21, from https://www.cambridge.org/core/journals/political-analysis/article/crosslingual-classification-of-political-texts-using-multilingual-sentence-embeddings/30689C8798F097EEBA514ABE4891A71B (Publisher: Cambridge University Press) doi: 10.1017/pan.2022.29

Licht, H., & Lind, F. (2023, September). Going cross-lingual: A guide to multilingual text analysis. *Computational Communication Research*, *5*(2). Retrieved 2024-01-05, from https://computationalcommunication.org/ccr/article/view/201 (Number: 2) doi: 10.5117/CCR2023.2.2.LICH

Lind, F., Heidenreich, T., Kralj, C., & Boomgaarden, H. G. (2021, October). Greasing the wheels for comparative communication research: Supervised text classification for multilingual corpora. *Computational Communication Research*, *3*(3). Retrieved 2022-08-02, from https://www.aup-online.com/content/journals/10.5117/CCR2021.3.001.LIND (Publisher: Amsterdam University Press) doi: 10.5117/CCR2021.3.001.LIND

Liu, A., Swayamdipta, S., Smith, N. A., & Choi, Y. (2022, December). WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 6826–6847). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Retrieved 2023-11-14, from https://aclanthology.org/2022.findings-emnlp.508 doi: 10.18653/v1/2022.findings-emnlp.508

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., … Roberts, A. (2023, February). *The Flan Collection: Designing Data and Methods for Effective Instruction Tuning.* arXiv. Retrieved 2023-03-29, from http://arxiv.org/abs/2301.13688 (arXiv:2301.13688 [cs])

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., … Hooker, S. (2023, November). *The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI.* arXiv. Retrieved 2023-11-14, from http://arxiv.org/abs/2310.16787 (arXiv:2310.16787 [cs])

Lou, R., Zhang, K., & Yin, W. (2023, May). *Is Prompt All You Need? No. A Comprehensive and Broader View of Instruction Learning.* arXiv. Retrieved 2023-11-19, from http://arxiv.org/abs/2303.10475 (arXiv:2303.10475 [cs])

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, *23*(2), 254–277. Retrieved 2021-12-19, from https://www.cambridge.org/core/product/identifier/S1047198700011736/type/journal_article doi: 10.1093/pan/mpu019

Ma, T., Yao, J.-G., Lin, C.-Y., & Zhao, T. (2021). Issues with Entailment-based Zero-shot Text Classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 786–796). Online: Association for Computational Linguistics. Retrieved 2022-01-07, from https://aclanthology.org/2021.acl-short.99 doi: 10.18653/v1/2021.acl-short.99

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning Word Vectors for Sentiment Analysis. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142–150). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved 2024-03-22, from https://aclanthology.org/P11-1015

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014, April). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782–796. Retrieved 2024-03-22, from https://doi.org/10.1002/asi.23062 doi: 10.1002/asi.23062

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021, May). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the*

AAAI Conference on Artificial Intelligence, *35*(17), 14867–14875. Retrieved 2024-03-22, from https://ojs.aaai.org/index.php/AAAI/article/view/17745 doi: 10.1609/aaai.v35i17.17745

McAuley, J., & Leskovec, J. (2013, October). Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems* (pp. 165–172). Hong Kong China: ACM. Retrieved 2024-03-22, from https://dl.acm.org/doi/10.1145/2507157.2507163 doi: 10.1145/2507157.2507163

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021, July). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 115:1–115:35. Retrieved 2022-08-03, from https://doi.org/10.1145/3457607 doi: 10.1145/3457607

Merz, N., Regel, S., & Lewandowski, J. (2016, April). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, *3*(2), 2053168016643346. Retrieved 2023-04-24, from https://doi.org/10.1177/2053168016643346 (Publisher: SAGE Publications Ltd) doi: 10.1177/2053168016643346

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., … Scialom, T. (2023, February). *Augmented Language Models: a Survey.* arXiv. Retrieved 2023-02-20, from http://arxiv.org/abs/2302.07842 (arXiv:2302.07842 [cs])

Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., & Scialom, T. (2023, November). *GAIA: a benchmark for General AI Assistants.* arXiv. Retrieved 2023-12-30, from http://arxiv.org/abs/2311.12983 (arXiv:2311.12983 [cs]) doi: 10.48550/arXiv.2311.12983

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems* (Vol. 26). Curran Associates, Inc. Retrieved 2022-02-26, from https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

Miller, B., Linder, F., & Mebane, W. R. (2020, October). Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches. *Political Analysis*, *28*(4), 532–551. Retrieved 2022-02-02, from https://www.cambridge.org/core/journals/

political-analysis/article/abs/active-learning-approaches-for-labeling-text-review-and-assessment-of-the-performance-of-active-learning-approaches/CF6AAF05F465D5BC688A9548433123C1#access-block (Publisher: Cambridge University Press) doi: 10.1017/pan.2020.4

Montani, I., Honnibal, M., Honnibal, M., Van Landeghem, S., Boyd, A., Peters, H., … Baumgartner, P. (2022, March). *explosion/spaCy: v3.2.4.* Zenodo. Retrieved 2022-09-21, from https://zenodo.org/record/6394862 doi: 10.5281/ZENODO.6394862

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., … Schulman, J. (2022, June). *WebGPT: Browser-assisted question-answering with human feedback.* arXiv. Retrieved 2023-04-01, from http://arxiv.org/abs/2112.09332 (arXiv:2112.09332 [cs]) doi: 10.48550/arXiv.2112.09332

Newton, P. E. (2012, January). Clarifying the Consensus Definition of Validity. *Measurement: Interdisciplinary Research and Perspectives*, *10*(1-2), 1–29. Retrieved 2022-11-16, from https://doi.org/10.1080/15366367.2012.669666 (Publisher: Routledge _eprint: https://doi.org/10.1080/15366367.2012.669666) doi: 10.1080/15366367.2012.669666

Newton, P. E., & Baird, J.-A. (2016, April). The great validity debate. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 173–177. Retrieved 2022-11-16, from https://doi.org/10.1080/0969594X.2016.1172871 (Publisher: Routledge _eprint: https://doi.org/10.1080/0969594X.2016.1172871) doi: 10.1080/0969594X.2016.1172871

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020, May). Adversarial NLI: A New Benchmark for Natural Language Understanding. *arXiv:1910.14599 [cs]*. Retrieved 2021-12-24, from http://arxiv.org/abs/1910.14599 (arXiv: 1910.14599)

OpenAI. (2023a). *GPT4 System Card* (Tech. Rep.). Retrieved from https://cdn.openai.com/papers/gpt-4-system-card.pdf

OpenAI. (2023b, March). *GPT-4 Technical Report.* arXiv. Retrieved 2023-05-01, from http://arxiv.org/abs/2303.08774 (arXiv:2303.08774 [cs]) doi: 10.48550/arXiv.2303.08774

Osnabrügge, M., Ash, E., & Morelli, M. (2021, October). Cross-Domain Topic Classification for Political Texts. *Political Analysis*, 1–22. Retrieved 2021-12-23, from https://

www.cambridge.org/core/journals/political-analysis/article/crossdomain-topic-classification-for-political-texts/F074564984969CE168BCBCF5E7D931C8 (Publisher: Cambridge University Press) doi: 10.1017/pan.2021.37

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., … Lowe, R. (2022, March). *Training language models to follow instructions with human feedback.* arXiv. Retrieved 2022-12-21, from http://arxiv.org/abs/2203.02155 (arXiv:2203.02155 [cs])

Pan, S. J., & Yang, Q. (2010, October). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. Retrieved 2021-12-27, from http://ieeexplore.ieee.org/document/5288526/ doi: 10.1109/TKDE.2009.191

Pang, B., & Lee, L. (2005, June). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In K. Knight, H. T. Ng, & K. Oflazer (Eds.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 115–124). Ann Arbor, Michigan: Association for Computational Linguistics. Retrieved 2024-03-22, from https://aclanthology.org/P05-1015 doi: 10.3115/1219840.1219855

Parrish, A., Huang, W., Agha, O., Lee, S.-H., Nangia, N., Warstadt, A., … Bowman, S. R. (2021, April). Does Putting a Linguist in the Loop Improve NLU Data Collection? *arXiv:2104.07179 [cs]*. Retrieved 2021-11-10, from http://arxiv.org/abs/2104.07179 (arXiv: 2104.07179)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. Retrieved 2022-10-14, from http://jmlr.org/papers/v12/pedregosa11a.html

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved 2022-09-21, from https://aclanthology.org/D14-1162 doi: 10.3115/v1/D14-1162

Pessach, D., & Shmueli, E. (2022, February). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *55*(3), 51:1–51:44.

Retrieved 2022-08-03, from https://doi.org/10.1145/3494672 doi: 10.1145/3494672

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, June). Deep Contextualized Word Representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved 2023-11-18, from https://aclanthology.org/N18-1202 doi: 10.18653/v1/N18-1202

Peterson, A., & Spirling, A. (2018, January). Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems. *Political Analysis*, *26*(1), 120–128. Retrieved 2022-02-05, from http://www.cambridge.org/core/journals/political-analysis/article/classification-accuracy-as-a-substantive-quantity-of-interest-measuring-polarization-in-westminster-systems/45746D999CFCD1CB43E362392D7B2FB4 (Publisher: Cambridge University Press) doi: 10.1017/pan.2017.39

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., … Gurevych, I. (2020). AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 46–54). Online: Association for Computational Linguistics. Retrieved 2022-11-17, from https://www.aclweb.org/anthology/2020.emnlp-demos.7 doi: 10.18653/v1/2020.emnlp-demos.7

Press, O., Smith, N. A., & Lewis, M. (2022, April). *Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation.* arXiv. Retrieved 2023-11-14, from http://arxiv.org/abs/2108.12409 (arXiv:2108.12409 [cs])

Project, P. A. (2014). *US Supreme Court Cases.* Retrieved 2022-03-13, from https://www.comparativeagendas.net/datasets_codebooks

Project, P. A. (2015). *US State of the Union Speeches.* Retrieved 2022-03-13, from https://www.comparativeagendas.net/datasets_codebooks

Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., … Sun, M. (2023, April). *Tool Learning with Foundation Models.* arXiv.

Retrieved 2023-04-20, from http://arxiv.org/abs/2304.08354 (arXiv:2304.08354 [cs]) doi: 10.48550/arXiv.2304.08354

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training.* Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019, February). *Language Models are Unsupervised Multitask Learners.* Retrieved from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2020, July). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* arXiv. Retrieved 2022-12-15, from http://arxiv.org/abs/1910.10683 (arXiv:1910.10683 [cs, stat])

Raman, M., Maini, P., Kolter, J. Z., Lipton, Z. C., & Pruthi, D. (2023, March). *Model-tuning Via Prompts Makes NLP Models Adversarially Robust.* arXiv. Retrieved 2023-03-26, from http://arxiv.org/abs/2303.07320 (arXiv:2303.07320 [cs])

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019, July). Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5370–5381). Florence, Italy: Association for Computational Linguistics. Retrieved 2024-03-22, from https://aclanthology.org/P19-1534 doi: 10.18653/v1/P19-1534

Reimers, N., & Gurevych, I. (2019, August). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* arXiv. Retrieved 2022-11-08, from http://arxiv.org/abs/1908.10084 (arXiv:1908.10084 [cs])

Reimers, N., & Gurevych, I. (2020, October). *Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation.* arXiv. Retrieved 2022-11-08, from http://arxiv.org/abs/2004.09813 (arXiv:2004.09813 [cs]) doi: 10.48550/arXiv.2004.09813

Rodman, E. (2020, January). A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. *Polit-*

ical Analysis, *28*(1), 87–111. Retrieved 2022-03-12, from http://www.cambridge.org/core/journals/political-analysis/article/timely-intervention-tracking-the-changing-meanings-of-political-concepts-with-word-vectors/DDF3B5833A12E673EEE24FBD9798679E (Publisher: Cambridge University Press) doi: 10.1017/pan.2019.23

Rodriguez, P. L., & Spirling, A. (2022, January). Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. *The Journal of Politics*, *84*(1), 101–115. Retrieved 2022-06-18, from https://www.journals.uchicago.edu/doi/full/10.1086/715162 (Publisher: The University of Chicago Press) doi: 10.1086/715162

Ruder, S. (2019). *Neural Transfer Learning for Natural Language Processing*. Ireland: National University of Ireland, Galway. Retrieved from https://ruder.io/thesis/neural_transfer_learning_for_nlp.pdf

Ruder, S., Vulić, I., & Søgaard, A. (2019, August). A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, *65*, 569–631. Retrieved 2022-11-07, from https://jair.org/index.php/jair/article/view/11640 doi: 10.1613/jair.1.11640

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020, February). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv. Retrieved 2024-01-13, from http://arxiv.org/abs/1910.01108 (arXiv:1910.01108 [cs]) doi: 10.48550/arXiv.1910.01108

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., … Rush, A. M. (2022, March). *Multitask Prompted Training Enables Zero-Shot Task Generalization*. arXiv. Retrieved 2022-12-21, from http://arxiv.org/abs/2110.08207 (arXiv:2110.08207 [cs])

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020, July). Social Bias Frames: Reasoning about Social and Power Implications of Language. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5477–5490). Online: Association for Computational Linguistics. Retrieved 2024-03-22, from https://aclanthology.org/2020.acl-main.486 doi: 10.18653/v1/2020.acl-main.486

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., & Chen, Y.-S. (2018, October). CARER: Contextualized Affect Representations for Emotion Recognition. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3687–3697). Brussels, Belgium: Association for Computational Linguistics. Retrieved 2024-03-22, from https://aclanthology.org/D18-1404 doi: 10.18653/v1/D18-1404

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., … Scialom, T. (2023, February). *Toolformer: Language Models Can Teach Themselves to Use Tools.* arXiv. Retrieved 2023-04-01, from http://arxiv.org/abs/2302.04761 (arXiv:2302.04761 [cs]) doi: 10.48550/arXiv.2302.04761

Schick, T., & Schütze, H. (2021a, April). Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 255–269). Online: Association for Computational Linguistics. Retrieved 2023-05-30, from https://aclanthology.org/2021.eacl-main.20 doi: 10.18653/v1/2021.eacl-main.20

Schick, T., & Schütze, H. (2021b, April). It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *arXiv:2009.07118 [cs]*. Retrieved 2021-12-24, from http://arxiv.org/abs/2009.07118 (arXiv: 2009.07118)

Schröder, C., & Niekler, A. (2020, August). *A Survey of Active Learning for Text Classification using Deep Neural Networks.* arXiv. Retrieved 2022-11-19, from http://arxiv.org/abs/2008.07267 (arXiv:2008.07267 [cs]) doi: 10.48550/arXiv.2008.07267

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015, May). Big Data, Digital Media, and Computational Social Science: Possibilities and Perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13. Retrieved 2024-01-04, from http://journals.sagepub.com/doi/10.1177/0002716215572084 doi: 10.1177/0002716215572084

Sileo, D. (2023, May). *tasksource: A Dataset Harmonization Framework for Streamlined NLP Multi-Task Learning and Evaluation.* arXiv. Retrieved 2023-11-14, from http://arxiv.org/abs/2301.05948 (arXiv:2301.05948 [cs]) doi: 10.48550/arXiv.2301.05948

Soups, R. (2015, February). *Yelp Dataset Challenge is Doubling Up!* Re-

trieved 2024-03-22, from https://engineeringblog.yelp.com/2015/02/yelp-dataset-challenge-is-doubling-up.html

Stewart, B. M., & Zhukov, Y. M. (2009, June). Use of force and civil–military relations in Russia: an automated content analysis. *Small Wars & Insurgencies*, *20*(2), 319–343. Retrieved 2022-02-05, from http://www.tandfonline.com/doi/abs/10.1080/09592310902975455 doi: 10.1080/09592310902975455

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2023, November). *RoFormer: Enhanced Transformer with Rotary Position Embedding.* arXiv. Retrieved 2023-11-14, from http://arxiv.org/abs/2104.09864 (arXiv:2104.09864 [cs]) doi: 10.48550/arXiv.2104.09864

Sun, Y., Zheng, Y., Hao, C., & Qiu, H. (2022, October). *NSP-BERT: A Prompt-based Few-Shot Learner Through an Original Pre-training Task–Next Sentence Prediction.* arXiv. Retrieved 2023-05-24, from http://arxiv.org/abs/2109.03564 (arXiv:2109.03564 [cs])

Taori, R., Gulrajani, I., Tianyi, Z., Dubois, Y., Xuechen, L., Guestrin, C., … B. Hashimoto, T. (2023). *Alpaca: A Strong, Replicable Instruction-Following Model.* Retrieved 2023-11-14, from https://github.com/tatsu-lab/stanford_alpaca

Terechshenko, Z., Linder, F., Padmakumar, V., Liu, M., Nagler, J., Tucker, J. A., & Bonneau, R. (2020, October). *A Comparison of Methods in Political Science Text Classification: Transfer Learning Language Models for Politics* (SSRN Scholarly Paper No. ID 3724644). Rochester, NY: Social Science Research Network. Retrieved 2021-12-23, from https://papers.ssrn.com/abstract=3724644 doi: 10.2139/ssrn.3724644

Theocharis, Y., & Jungherr, A. (2021, March). Computational Social Science and the Study of Political Communication. *Political Communication*, *38*(1-2), 1–22. Retrieved 2022-12-22, from https://doi.org/10.1080/10584609.2020.1833121 (Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2020.1833121) doi: 10.1080/10584609.2020.1833121

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018, December). *FEVER: a large-scale dataset for Fact Extraction and VERification.* arXiv. Retrieved 2024-03-22, from http://arxiv.org/abs/1803.05355 (arXiv:1803.05355 [cs]) doi: 10.48550/arXiv.1803.05355

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei,

Y., … Scialom, T. (2023, July). *Llama 2: Open Foundation and Fine-Tuned Chat Models.* arXiv. Retrieved 2023-11-14, from http://arxiv.org/abs/2307.09288 (arXiv:2307.09288 [cs]) doi: 10.48550/arXiv.2307.09288

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., … Wolf, T. (2023, October). *Zephyr: Direct Distillation of LM Alignment.* arXiv. Retrieved 2023-11-14, from http://arxiv.org/abs/2310.16944 (arXiv:2310.16944 [cs]) doi: 10.48550/arXiv.2310.16944

Unknown. (2024, February). *yahoo_answers_topics · Datasets at Hugging Face.* Retrieved 2024-03-22, from https://huggingface.co/datasets/yahoo_answers_topics

Van Atteveldt, W., & Peng, T.-Q. (2018, April). When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, *12*(2-3), 81–92. Retrieved 2021-12-21, from https://doi.org/10.1080/19312458.2018.1458084 (Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2018.1458084) doi: 10.1080/19312458.2018.1458084

Van Atteveldt, W., & Peng, T.-Q. (2021). *Computational Methods for Communication Science*. Routledge. (Google-Books-ID: _PsTEAAAQBAJ)

Van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019, September). Toward Open Computational Communication Science: A Practical Road Map for Reusable Data and Code. *International Journal of Communication*, *13*(0), 20. Retrieved 2022-01-16, from https://ijoc.org/index.php/ijoc/article/view/10631 (Number: 0)

Van Atteveldt, W., Trilling, D., & Calderon, C. A. (2022). *Computational Analysis of Communication* (1st edition ed.). Hoboken, NJ: Wiley-Blackwell. Retrieved from https://cssbook.net/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. Retrieved 2023-05-30, from https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wallach, H. (2018, February). Computational social science computer

science + social data. *Communications of the ACM*, *61*(3), 42–44. Retrieved 2024-01-04, from https://dl.acm.org/doi/10.1145/3132698 doi: 10.1145/3132698

Wang, S., Fang, H., Khabsa, M., Mao, H., & Ma, H. (2021, April). Entailment as Few-Shot Learner. *arXiv:2104.14690 [cs]*. Retrieved 2021-11-05, from http://arxiv.org/abs/2104.14690 (arXiv: 2104.14690)

Wang, X., Wang, H., & Yang, D. (2022, May). *Measure and Improve Robustness in NLP Models: A Survey.* arXiv. Retrieved 2023-03-26, from http://arxiv.org/abs/2112.08313 (arXiv:2112.08313 [cs])

Webson, A., & Pavlick, E. (2022, July). Do Prompt-Based Models Really Understand the Meaning of Their Prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2300–2344). Seattle, United States: Association for Computational Linguistics. Retrieved 2023-04-30, from https://aclanthology.org/2022.naacl-main.167 doi: 10.18653/v1/2022.naacl-main.167

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., … Le, Q. V. (2022, February). *Finetuned Language Models Are Zero-Shot Learners.* arXiv. Retrieved 2022-12-21, from http://arxiv.org/abs/2109.01652 (arXiv:2109.01652 [cs])

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2023, January). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* arXiv. Retrieved 2023-11-25, from http://arxiv.org/abs/2201.11903 (arXiv:2201.11903 [cs]) doi: 10.48550/arXiv.2201.11903

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., … Gabriel, I. (2021, December). *Ethical and social risks of harm from Language Models.* arXiv. Retrieved 2023-11-19, from http://arxiv.org/abs/2112.04359 (arXiv:2112.04359 [cs])

Widmann, T., & Wich, M. (2022, June). Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text. *Political Analysis*, 1–16. Retrieved 2022-07-06, from http://www.cambridge.org/core/journals/political-analysis/article/creating-and-comparing-dictionary-word-embedding-and-transformerbased-models

-to-measure-discrete-emotions-in-german-political-text/2DA41C0F09DE1CA600B3DCC647302637 (Publisher: Cambridge University Press) doi: 10.1017/pan.2022.15

Wilkerson, J., & Casas, A. (2017, May). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science*, *20*(1), 529–544. Retrieved 2021-12-19, from https://www.annualreviews.org/doi/10.1146/annurev-polisci-052615-025542 doi: 10.1146/annurev-polisci-052615-025542

Williams, A., Nangia, N., & Bowman, S. (2018, June). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1112–1122). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved 2023-05-30, from https://aclanthology.org/N18-1101 doi: 10.18653/v1/N18-1101

Windsor, L. C., Cupit, J. G., & Windsor, A. J. (2019). Automated content analysis across six languages. *PLOS ONE*, 14. doi: https://doi.org/10.1371/journal.pone.0224425

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. (2020, October). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved 2023-05-30, from https://aclanthology.org/2020.emnlp-demos.6 doi: 10.18653/v1/2020.emnlp-demos.6

Wu, P. Y., Nagler, J., Tucker, J. A., & Messing, S. (2023, September). *Large Language Models Can Be Used to Estimate the Latent Positions of Politicians.* arXiv. Retrieved 2024-01-13, from http://arxiv.org/abs/2303.12057 (arXiv:2303.12057 [cs]) doi: 10.48550/arXiv.2303.12057

Xia, M., Artetxe, M., Du, J., Chen, D., & Stoyanov, V. (2022, December). Prompting ELECTRA: Few-Shot Learning with Discriminative Pre-Trained Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 11351–11361). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Retrieved 2023-06-06, from https://aclanthology.org/2022.emnlp-main.780

Xu, H., Lin, Z., Zhou, J., Zheng, Y., & Yang, Z. (2023, July). A Universal Discriminator for Zero-Shot Generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 10559–10575). Toronto, Canada: Association for Computational Linguistics. Retrieved 2023-08-02, from https://aclanthology.org/2023.acl-long.589

Xu, P., Zhu, X., & Clifton, D. A. (2023, May). *Multimodal Learning with Transformers: A Survey.* arXiv. Retrieved 2023-11-25, from http://arxiv.org/abs/2206.06488 (arXiv:2206.06488 [cs])

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., … Hu, X. (2023, April). *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond.* arXiv. Retrieved 2023-04-30, from http://arxiv.org/abs/2304.13712 (arXiv:2304.13712 [cs]) doi: 10.48550/arXiv.2304.13712

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023, March). *ReAct: Synergizing Reasoning and Acting in Language Models.* arXiv. Retrieved 2023-04-01, from http://arxiv.org/abs/2210.03629 (arXiv:2210.03629 [cs])

Yao, Y., Dong, B., Zhang, A., Zhang, Z., Xie, R., Liu, Z., … Wang, J. (2022, May). Prompt Tuning for Discriminative Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 3468–3473). Dublin, Ireland: Association for Computational Linguistics. Retrieved 2022-07-29, from https://aclanthology.org/2022.findings-acl.273 doi: 10.18653/v1/2022.findings-acl.273

Yin, W., Hay, J., & Roth, D. (2019, November). Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3914–3923). Hong Kong, China: Association for Computational Linguistics. Retrieved 2023-05-30, from https://aclanthology.org/D19-1404 doi: 10.18653/v1/D19-1404

Yin, W., Radev, D., & Xiong, C. (2021, June). DocNLI: A Large-scale Dataset for Document-level Natural Language Inference. *arXiv:2106.09449 [cs].* Retrieved 2021-12-24, from http://arxiv.org/abs/2106.09449 (arXiv: 2106.09449)

Yin, W., Rajani, N. F., Radev, D., Socher, R., & Xiong, C. (2020, October). Universal Natural Language Processing with Lim-

ited Annotations: Try Few-shot Textual Entailment as a Start. *arXiv:2010.02584 [cs]*. Retrieved 2021-12-24, from http://arxiv.org/abs/2010.02584 (arXiv: 2010.02584)

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc. Retrieved 2024-03-22, from https://papers.nips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., … Stoica, I. (2023, October). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.* arXiv. Retrieved 2023-11-19, from http://arxiv.org/abs/2306.05685 (arXiv:2306.05685 [cs]) doi: 10.48550/arXiv.2306.05685

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., … Levy, O. (2023, May). *LIMA: Less Is More for Alignment.* arXiv. Retrieved 2023-05-22, from http://arxiv.org/abs/2305.11206 (arXiv:2305.11206 [cs])

Zobel, M., & Lehmann, P. (2018, November). Positions and saliency of immigration in party manifestos: A novel dataset using crowd coding. *European Journal of Political Research*, *57*(4), 1056–1083. Retrieved 2021-12-21, from https://onlinelibrary.wiley.com/doi/10.1111/1475-6765.12266 doi: 10.1111/1475-6765.12266