Fairness in Transfer Learning for Natural Language Processing

Seraphina Goldfarb-Tarrant

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2024

Abstract

Natural Language Processing (NLP) systems have come to permeate so many areas of daily life that it is difficult to live a day without having one or many experiences mediated by an NLP system. These systems bring with them many promises: more accessible information in more languages, real-time content moderation, more data-driven decision making, intuitive access to information via Question Answering and chat interfaces. But there is a dark side to these promises, for the past decade of research has shown that NLP systems can contain social biases and deploying them can incur serious social costs. Each of these promises has been found to have unintended consequences: racially charged errors and rampant gender stereotyping in language translation, censorship of minority voices and dialects, Human Resource systems that discriminate based on demographic data, a proliferation of toxic generated text and misinformation, and many subtler issues.

Yet despite these consequences, and the proliferation of bias research attempting to correct them, NLP systems have not improved very much. There are a few reasons for this. First, measuring bias is difficult; there are not standardised methods of measurement, and much research relies on one-off methods that are often insufficiently careful and thoroughly tested. Thus many works have contradictory results that cannot be reconciled, because of minor differences or assumptions in their metrics. Without thorough testing, these metrics can even mislead and give the illusion of progress. Second, much research adopts an overly simplistic view of the causes and mediators of bias in a system. NLP systems have multiple components and stages of training, and many works test fairness at only one stage. They do not study how different parts of the system interact, and how fairness changes during this process. So it is unclear whether these isolated results will hold in the full complex system. Here, we address both of these shortcomings. We conduct a detailed analysis of fairness metrics applied to upstream language models (models that will be used in a downstream task in transfer learning). We find that a) the most commonly used upstream fairness metric is not predictive of downstream fairness, such that it should not be used but that b) information theoretic probing is a good alternative to these existing fairness metrics, as we find it is both predictive of downstream bias and robust to different modelling choices. We then use our findings to track how unfairness, having entered a system, persists and travels throughout it. We track how fairness issues travel between tasks (from language modelling to classification) in monolingual transfer learning, and between languages, in multilingual transfer learning. We find that multilingual transfer learning often exacerbates fairness problems and should be used with care, whereas monolingual transfer learning generally improves fairness. Finally, we track how fairness travels between source documents and retrieved answers to questions, in fact-based generative systems. Here we find that, though retrieval systems strongly represent demographic data such as gender, bias in retrieval question answering benchmarks does not come from the model representations, but from the queries or the corpora. We reach all of our findings only by looking at the entire transfer learning system as a whole, and we hope that this encourages other researchers to do the same. We hope that our results can guide future fairness research to be more consistent between works, better predictive of real world fairness outcomes, and better able to prevent unfairness from propagating between different parts of a system.

Lay Summary

Natural Language Processing describes any AI system that deals with human language. Today NLP systems are part of every person's daily life; visibly as phone voice assistants and automatic YouTube captioning, and invisibly, when that person applies for a job or when all the data they put online is analysed for content moderation, marketing, opinion polling, and more. **Transfer Learning** describes systems that are multi-part, which almost all systems are today, because they are built from one big language model, like ChatGPT, Llama, Mistral, or Cohere, which is later *fine-tuned* to a particular task, such as customer service or resume processing. Fine-tuning just means that some data in that domain, usually expensive, often proprietary, is used after the model is trained in order to make the language model less general purpose and better at that type of data. This system described so far is Transfer Learning. In many of today's systems, the language model is then connected to a RAG (Retrieval Augmented Generation) component, where the model can search a database of documents: Wikipedia, confidential medical records, all the PDFs that a PhD student has downloaded over the course of their degree[1]. The premise of the research in this thesis is that you have to know how the parts of these systems interact, in order to judge whether a system is **fair**. This means the interaction between all of: the original big language model, the data you fine-tune on, anything else you make at the fine-tuning stage, the retrieval system, the corpus it retrieves from. A system that is **fair** is one that doesn't propagate social biases and stereotypes, and doesn't screw over minority groups in society. To be precise to the definition of fairness, it shouldn't screw over *any* group, even straight white men, but the minority ones tend to be the ones we worry about since they tend to be most screwed over, with some exceptions.

In this work, we discover a couple of things about how the parts of an NLP system interrelate, with regard to fairness. First, you cannot measure fairness of just the language model in isolation and know whether your model will be fair or unfair when it's deployed in an application later. You can get an indication as to the model's *potential* for unfairness, but that is all. This idea of *potential* can be understood by analogy to genetics—we can measure whether a person is more or less likely to develop cancer, but whether they do or not over the course of their life depends on whether they work in copper mines and what they eat, how much they exercise and what pollutants are in the air where they live. Some people with a high propensity may never develop it, and

[1] Roughly 1227 papers, in my case.

some people with no propensity at all still will if they live next to Chernobyl. With language models, we can measure how much potential there is for unfairness, but we can't know what this really means without the environmental factors: the fine-tuning data, the RAG system, and even the cultural context of deployment, which determines what demographic groups are considered minorities. We can make predictions about how likely the model will be to become unfair, if it is fine-tuned on skewed data that doesn't have positive examples of good resumes for people who aren't white, or doesn't have high quality RAG data for people who are not male (this is, incidentally, true of Wikipedia (Sun and Peng, 2021)). So we can improve, or mitigate this *potential* in a language model. But in real world applications, where we need *know* that people are not screwed over with some degree of certainty, we need to test the final system.

Most language models today are also **multilingual**. If you type something into ChatGPT or Cohere's model in English, it can retrieve information from documents in Turkish and summarise them for you in English. If you type something in Korean, it will respond in Korean. In this work, we discovered that data in one language can influence fairness behaviour in *another* language. An NLP system that handles Japanese can become more sexist and racist when you add data from English, even though the data you're adding is not in Japanese, and even though the sexist and racist stereotypes in English data are not the same as those in native Japanese.

Overall, this work shows that you cannot assume that the addition of new data—from fine-tuning, for other languages, from RAG—does not change the fairness properties of an original model. So fairness cannot just be the domain of the tech giants and gold-plated start-ups, of Google, Meta, Anthropic, Cohere, or whatever companies are producing the new hot model next year. Is has to be a collaboration between the people training the big models, the people deploying systems in the real world, and, ideally, the users.

Acknowledgements

People say that PhDs are the most solitary time of your life. In some ways, I've found that to be true. But I very much did not do this alone.

The inkling in my mind that I wanted to leave my career and go study NLP began roughly ten years ago, and I would not be here today without the herculean efforts of some, and the small graces of many. I want to thank first Somusa Ratanarak, who encouraged me to leave Google and pursue my passions when all my colleagues thought leaving a stable job was utterly insane. She's been with me every step of the way, with a lot of cleverness and care, and a little bit of 愛の鞭, and I am a much better person because of her. I want to thank my parents, all of them, Mom#1, Mom#2, Dad, Tyger, and Rachel, who also didn't think I was crazy. Thank you for seeing and celebrating my accomplishments, with flowers and lemons and jam, even when I didn't stop to appreciate them because I was busy running for the next goalpost. Thank you for being proud of me.

There were still others whose help I needed even to reach the start of my PhD. I want to thank Ozan Mindek, for giving me my first chance at NLP on a 20% project. Then Emily Bender, for walking her talk and creating an MSc program that was *truly* diverse in practice. She took me with my undergraduate in Ancient Greek, and my now friends Brian from the Navy and Lonny working on Yupiq preservation, along with the usual CS suspects. I aspire to be that conscientious in organisations that I lead, and I can see already how challenging it will be to live up to. I also need to thank that same diverse cohort of Brian, Lonny, Catharine, Genevieve, Amandalynne, Chris, and Ben for spending Saturday mornings helping me practice interview for my PhD. And I aspire also to be like two more women who helped me: Fei Xia, who can make absolutely any question you ask her interesting, no matter how silly it is, and who showed me the beauty and elegance of statistical machine learning. And Nanyun Peng, for taking on a green MSc student and teaching me to write papers and staying up late with me, and showing me that I absolutely loved doing research.

Now we have reached the start of my PhD. Here I want to thank both my advisors. Björn, for taking me on partway through and bringing his energy and enthusiasm, at a point in the pandemic when my own energy was flagging; our Meadows walks kept me going. Adam, for accepting me, trusting whatever direction I ran off in, and immediately being for me a mentor who cares deeply about both good science and about people. Adam has changed the way I do science and the way I see the world; I am proud to be in his intellectual lineage.

So many other people made my PhD possible, by keeping me afloat along the way, in a period in which very many things went wrong in my life and in the world. I owe to them, and to those moments, both a piece of this research and of who I am.

Late nights walking the City of London with Sameer Bansal, whose dark humour and enormous generosity lightens all loads. Hikes, bothies, wine and cheese nights with Kate McCurdy, who would always make me laugh and teach me something marvelous about humans or language, and who became a better friend and colleague than I could've ever wished for. Always-inadvisably-late nights by the fire with Henry Conklin, with whom I get my most interesting NLP inspiration and also with whom I can be my whole self, unfiltered. All my adventures in the beautiful wilds with Ida Szubert, who holds all of my soul and of my delight in life, and with whom it is so easy to share space, and somehow never a chore to work late, or do anything hard, because I am with her. Pints at Sandy Bells with Ramon Sanabria, who sees the cheerful beauty in even quite shitty situations and shows me how to do this too. Dinners and walks with Yevgen Matusevych, who helped me when I needed it and scolded me for not asking. Coffees with Tom Hosking, who has a knack for re-energising me about the whole field of NLP over coffee, just when I've been banging my head against a problem so long it's become tedious. Writing accountability meetings at the Burn with Tom Sherborne, who was instrumental in getting my nightmare multilingual work onto a page, and has been a companion since. Brewing the worst beer I have ever made with Sander Bijl de Vroe and drinking it anyway. Walks down the Innocent Railway with Siddarth Narayanaswamy, reminding me that I wished I'd learnt first principles better, but also that this wish is why I became a scientist in the first place. Whisky nights with Jasmijn Bastings, as well as our long chat on my 15-hour solo drive South from NAACL after I caught Covid-19 there. She was always so easygoing, and foolish/caring enough to come to all my panels and talks, without me asking.

All of my direct collaborators also kept me going through a PhD that was much, much lonelier than I could have ever anticipated. Meetings with my collaborators Hadas Orgad and Yonatan Belinkov, which I would look forward to from the start of every week and would keep me energised until the end of it. Adam and Sharon's Agora research group in ILCC provided feedback, structure, and scholarly companionship. Diego Marcheggiani and Roi Blanco helped me manage my first enormous multilingual project. Patrick Lewis and Pedro Rodriguez exemplified a combination of rigor and curiosity in their science, as well as conscience, that I admire and hope to live up to.

I am thankful for the small graces (and some large graces) far beyond the research sphere also. To James and Deena Owers-Bardsley for my intro to cycling in Scotland and for the cocktail kits we left on each others' doorsteps when we couldn't see other people. To David Halliwell and Narma Gebruk, for celebrating my *intended* submission with me last May, and keeping the gorse-flame alight the next nine months and then coming out with me to celebrate again. To the entire Beltane family — especially my performance groups Veles, Goblin Fire Arch, Goblin Bower, The Summer King, and Obsidian — who gave me a haven away from academia which at times I desperately needed. To Alison Stewart, for taking me in practically as a family member after that one coffee in George Square. To Sam Roots and Ruari Cathmoir, for so easily and comfortably becoming chosen family. To Ellen Mears, who has been my companion in doing the things that I want to do for *me*, rather than spending all my time on my sometimes heavy responsibilities. To Guru Khalsa, for every one of those calls from the road. To Ivan Ivanov, who was accidentally stuck with me during the pandemic, and who I hope to get stuck with many more times in my life. To James Hartley, who is one of the few people in the world I feel I can lean on, and who has always provided me with so much love it keeps me warm from another hemisphere. To Ezra Baydur, for drunk-chat-chess and seeing the best in me. To Craig Innes, who has introduced me to a version of myself that I didn't know was there before, and that I am so glad to have gotten to know; who reliably has something uplifting to say about my skills, to prop me up, before I give a scary public appearance. To Tasuku Koda, Narumi Ota, and Makoto Takashima, for showing me 本音 and keeping a place on the other side of the world that feels like home. To Alice in Slumberland, my Burning Man family, who I saw only once during my PhD but who taught me that I could express myself any way I wanted, and that I could even redefine the world in which I lived. I'm trying to do a little bit of that redefinition in here, with this work.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

My specific contributions are as follows for each chapter:

Chapter 3: I designed the research agenda: I envisioned the research question and wrote a research plan document with methods, goals, metrics, and a literature review. I recruited and supervised three MSc students who implemented pipelines for three different systems and did initial investigations into the correlation between intrinsic and extrinsic metrics for their MSc theses. I gathered these pipelines together, extended them, ran experiments, and wrote and presented the paper, with the help of Adam.

Chapter 4: I was second author on this paper, assisting the first author Hadas Orgad who proposed this extension of the work in Chapter 3. I implemented some of the metrics on intrinsic analysis of language model representations, implemented the additional extrinsic fairness metrics (which are now open-sourced), and co-wrote the paper with Hadas Orgad and Yonatan Belinkov.

Chapters 5 and 6: I did this project almost entirely on my own save for the writing, which my supervisors Adam and Björn assisted with greatly. I designed the research question and outlined the project, developed and programmed the experiment framework, found training data, created evaluation data (with the help of native speakers of each language) and wrote up the results (with the help of my supervisors). I received regular weekly consultation from Diego Marcheggiani, Roi Blanco, and Lluis Marquez at Amazon Barcelona for the first of the two projects.

Chapter 7: I designed the research question in collaboration with Patrick Lewis. I made the research project plan, chose datasets and interpretability methods, wrote all pipeline code and ran all experiments, and finally wrote up the findings into a paper, with help from Pedro Rodriguez.

(*Seraphina Goldfarb-Tarrant*)

Contents

Chapter 1

Introduction

In the past decade, Natural Language Processing (NLP) systems have come to saturate everyday life. NLP has expanded from being used to translate webpages and recommend new videos to having inescapable reach. It is now also used to moderate social media content (Winchcomb, 2019), where all posts are filtered through an NLP system that judges them as hateful/not-hateful, acceptable/not-acceptable, and either removes or suppresses the sharing of posts that fail this check. NLP is used to generate answers to any user questions about any topic (Rajpurkar et al., 2016, 2018), by sifting through millions of documents and determining which ones are relevant and worth knowing about and presenting those, discarding others. It is used to track public opinion about products or politicians (Nissim et al., 2020), by analysing the sentiment of all of the information said about them online. NLP is used to sort and filter resumes for potential new workers (Parasurama and Sedoc, 2022), by comparing each new resume to previous successful hires for a job, and flagging which ones to send to a phone-screen and which ones to reject. It is also used to later fire those same workers (Kelly, 2023), by reading data about their productivity and predicting who shouldn't make the cut. These examples are a small set of the myriad applications that use NLP today, chosen as the areas the are directly relevant to the research in this thesis. There are so many more; most people in the UK and US come into contact with NLP or AI multiple times daily, though many of them are not aware of it (Kennedy et al., 2023).

This increase in scope and usage of NLP systems comes with many promises of efficiency, cost reduction, and even social good. For all of the uses above, there are bright promises. NLP for content moderation on social media can reduce hatespeech and aggression online, which has reached a volume and velocity that is completely unmanageable for human moderators. When left unchecked, it is linked to amplification of violence in the real world. NLP for retrieval and question answering can enable greater and easier access to information, a necessary step in searching and organising the vast quantities of digital information and democratising information access. NLP as sentiment analysis of public opinion can enable direct and inexpensive democratic feedback for companies or policies; direct feedback that might otherwise be too logistically challenging and expensive to gather. NLP in hiring systems could enable processing more applications, which could give a broader segment of the population a chance, and create systems that are less dependent on 'who you know' and on the instincts of a few HR representatives tasked with reading through a resume slush pile.

But this territorial expansion introduces many new harms that diminish these promises. The often referenced promise of mathematical objectivity—freeing us from human subjectivity, inconsistency, and biases—has proven to be mythical. At best, NLP systems learn and propagate these same biases, but with a veneer of objectivity that fosters over-reliance (O'Neil, 2016) and reduces accountability and recourse when data is incorrect and decisions go wrong.

A large and growing body of work analysing NLP systems has shown that they do not behave similarly and work equally well for different genders, races, nationalities, and other demographic groups. This disparate performance across demographics is the standard definition of **fairness**, which we use throughout this thesis: these systems are not **fair**. So given that NLP systems, and the data, models, and optimisation and evaluation metrics they are composed of are *not* inherently fair, we must analyse the ways they are not, so we know what to expect and can mitigate where possible. When NLP systems are not fair, companies and organisations using them (and people subjected to their outputs) are worse off than before automation systems, since this flawed system has now been scaled. An individual Human Resources manager may have flaws and biases, but they work for only one or a few companies and have time to read only so many resumes in a day. Some resumes will be sent to a different person, who may have different biases, preventing the inequities of the first person's views from being complete and consistent over a wide swath of potential jobs. When a flawed and biased NLP HR system is scaled, it does not sleep, get tired, or clock-off and can process as many thousands of resumes as time and compute allows. The same system is used by many companies. The very variability of human behaviour, and the inconsistencies in human decisionmaking that are often considered undesirable, limit the possible scope of each individual's (or even each company or organisation's) biases. This lack of scaling of humans is an accidental safeguard. An NLP system, in contrast, replicates the same biases to an unlimited extent, and whatever unfortunate minorities it is biased against will experience more widespread discrimination. This is the situation of the present day, and sets the scene for this research.

The research world noticed this, eventually. Fairness problems in NLP started to become well-known in the NLP community in 2016, as NLP itself began to directly touch more lives and have more impact (Hovy and Spruit, 2016). Attention in the research world, and the public, has grown exponentially since[1]. By the time of writing, major conferences now have a dedicated track for fairness research[2], encourage papers to self-declare potential hazards[3], and have an ethics committee appointed to review potential fairness problems in any work[4]. Yet despite all the attention and effort expended on fairness in NLP, we as a community have made only such a small dent in known problems as to now be aware of the magnitude of still unaddressed fairness problems. Both discovering and addressing fairness problems in an NLP system remains extremely challenging.

There are a couple of reasons for this challenge, which have prevented the community from making a larger dent. One of the most salient ones is that NLP systems today are complex; they involve multiple stages of model training, as is the case with Transfer Learning (discussed and defined in §2.5). How to measure and mitigate unfairness in a multi-part NLP system is not clear, and systems are now always multi-part. How does a measurement or mitigation at one stage relate to the other stages? What can be trusted to hold across stages? This thesis attempts to take a step towards remedying this. It asserts, throughout each of the sections, that you cannot study just one part of an NLP system in isolation, without first understanding how it affects the other parts.

There is real difficulty in even defining unfairness, and a substantial percentage of fairness papers neglect to define it at all (Blodgett et al., 2020; Goldfarb-Tarrant et al., 2023). There are multiple ways that a system can be unfair. NLP systems are often not **allocationally fair**; they have different accuracies and rates of false positives and

[1] https://fairmlclass.github.io/1.html#/4

[2] https://aclrollingreview.org/cfp

[3] Section A2 in https://aclrollingreview.org/responsibleNLPresearch/, and section 1c in https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist

[4] https://www.aclweb.org/adminwiki/index.php?title=Formation_of_the_ACL_Ethics_Committee

false negatives for different demographics. A example such situation is when a toxicity detection system has much higher rates of false positives for text that is actually benign or beneficial but contains terms about race, religion, or sexual orientation. In such as case, a sentence like *I am a gay man* can be flagged as toxic and censored, as was the case with Google's toxicity detection system in 2018 (Dixon et al., 2018). NLP systems are also often not **representationally fair**; they reproduce and propagate negative stereotypes for minoritised demographics (Crawford, 2017). For example, prominent generation systems will disproportionately describe women as taking carer roles, and portray racial minorities as criminals (Sheng et al., 2019).

There is not even a consensus on how best to measure each type of unfairness. Most metrics used to measure fairness are ad-hoc and have not been standardised or analysed for **predictive validity**—their ability to predict actual fairness problems that will occur–or **concurrent validity**—their agreement with other metrics in use. If you cannot measure something, 'your knowledge is of a meagre and unsatisfactory kind' (Kelvin, 1891) and you cannot know whether any improvements you make actually worked. So we begin by making some progress towards assessing predictive and concurrent validity of fairness metrics in Part I.

Another challenge is that fairness issues can appear at almost any stage of building an NLP system (Suresh and Guttag, 2021), and as mentioned, the relationship between the stages is poorly understood. NLP papers commonly claim that 'model biases reflect biases in data they were trained on'[5] but this is such a gross oversimplification as to be both unhelpful and misleading. It glosses over questions such as: *how did the biases get into the data? Do imbalances in labels over different sensitive groups count, or do only stereotypes count?* And it glosses over all the other causes, of which there are seven high level kinds in Suresh and Guttag (2021), and some additional in other works (Mehrabi et al., 2021). I expect the prevalence of this statement is a way of shirking responsibility. If it is the data's fault for being biased, and society's fault for creating biased data, then it is not the fault of the engineer or company for creating a biased model. It's just the world we live in.

But it is not the world we live in. It is the world we are making. All choices in the process of training an NLP model can affect the resulting bias. A resume filtering

[5] This refrain is ubiquitous, and is apparently even the rationale that ChatGPT gives: *Yes, language models can have biases, because the training data reflects the biases present in society from which that data was collected.* as reported in https://news.mit.edu/2023/large-language-models-are-biased-can-logic-help-save-them-0303

system can be trained on data in which humans made racist or sexist decisions – say in the past they didn't hire non-men, or non-white, or only hired young people or people who went to certain schools. This is **historical** bias, and that bias in the training data will not only persist, but be amplified, an effect which is much less frequently discussed but is common (Zhao et al., 2017; Jia et al., 2020; Cabello et al., 2023; Hashimoto et al., 2018). Then this system, which is already dubiously 'only reflecting training data' will scale, with the authority of an objective AI system behind it. This happened with Amazon's attempt at an AI for Human Resources, which would not hire women at all because, historically, Amazon had not hired very many of them.[6] There are also **sampling**, aka **representation**, biases. For instance, a content moderation and toxicity detection system can be unfamiliar with non-prestige dialects and censor them incorrectly, as happened when tweets in African-American Vernacular English (AAVE) were incorrectly flagged as toxic speech (Sap et al., 2019). Even though AAVE is common in the *world*, it was not well-represented in data the model has seen. So even within dataset biases, there are multiple kinds with different reasons behind them. There are other sources beyond dataset biases. The above example of historical bias in hiring is actually an example of another type of bias as well, **measurement** bias. This NLP resume filtering system uses labelled data for supervised learning (as is quite common) where the labels are a proxy for the task that is to be learnt – e.g. *was previously hired based on this resume* is a proxy for *was suitable for the job*. That label can be a better or worse proxy for the desired task. This gap between the thing being measured and the unmeasurable quantity of interest is **measurement** bias. These are only a few of the many ways that unfairness can enter a system, selected as examples as they are the types that I spend the most time examining below. There are more subtle ways that can make mitigation even more challenging, which I discuss in Part 2.

An NLP system can contain one, many, or all of these sources of bias, and this bias can enter in via the data collection, dataset splits, learning objective, model architecture, model deployment choices (such as decoding hyperparameters or classifier thresholds). And most of these choices are now made *twice* or more. Current scale in NLP is driven by **transfer learning**, where a model is trained on high resource task(s) or language(s) (e.g. unstructured web crawl text) and then ported to a lower resourced one (e.g. any supervised task requiring labels, like sentiment analysis) – not necessarily objectively *low resource*, but relatively lower resourced, i.e. with less data than the dominant task

[6] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

or language being used for transfer. It was already difficult to pinpoint where biases enter a system, and with transfer learning most systems are composed of multiple sets of training data, multiple objectives, multiple measurements.

Transfer learning is now the dominant paradigm in NLP, but previous to the work in this thesis, fairness research considered only one of the two stages: the pre-training or the fine-tuning stage. If a language model that will later be used in our example resume filtering system (which we refer to as an **upstream model**) has been debiased with regard to gender, will the classifier on top of it (which we refer to as the **downstream model**) also be debiased, or not? If instead the classifier is debiased, is the language model also safe to use, or will bias then surface if the language model is used in another task, or directly without the classifier? We cannot answer these questions without studying the entire system and learning the relationship between upstream and downstream models. And without these answers, bias mitigation methods or measurements are at best ineffective, and at worst misleading. With these answers we can apply effective bias mitigation strategies at the correct stage of the system, and we will understand the contribution of transfer learning to fairness in NLP systems and be informed as to whether systems are becoming better or worse as they scale. This understanding is a pre-requisite to effective work in NLP bias, and yet before the work in this thesis, the field had little knowledge of it.

So here, in the below, we explore a previously yet unstudied area of NLP fairness; how unfairness, having entered a system, persists and travels throughout it.

We first focus, in Part I, on fairness measurement at different stages of transfer learning. No real research can be done without good measures, and we need an understanding of how measures of bias relate at different stages of transfer learning, since interventions are customarily applied at one stage. In Chapter 3, we study whether the most common **intrinsic** bias measurements–at the language model pre-training stage–are predictive of later downstream, or **extrinsic**, bias in two classification tasks in two languages. We find that they are not predictive, and that the widespread use of these measures has been leading to a false sense of progress in debiasing research. Most work was at the time done on only upstream models, and our work shows that we cannot tell whether debiasing efforts are propagating downstream. Our results show that more effort needs to be spent on measuring bias on the downstream task itself. Following this, in Chapter 4, we study the relationship of transfer learning measurements in the *reverse* direction. Here we ask how a pre-trained upstream language model changes when different debiasing methods are applied downstream. We find that a new metric, based on information theoretic probing (also known as minimum description length (MDL) probing) (Voita and Titov, 2020) can, when applied to the pre-trained language model, differentiate between different downstream bias levels, and different downstream debiasing techniques, and show which are more effective. We find that this measure is predictive of how robust debiasing of the pre-trained language model is, and whether the debiasing will remain if that model is then used in another task. These two results together imply that the **geometry** (cosine or other distance measures, previously used as upstream metrics) of concepts in language model representation space does not reliably predict downstream bias, but the **extractability** of concepts (as measured by information theoretic codelengths) is better predictor. In that work, we also are the first to use a wide suite of ten downstream fairness metrics that refer to slightly different notions of fairness. We find that though they tend to track together, if we had naively used a subset of them, based on what was most popular for certain datasets, we might have come to a different conclusion. Different metrics are suitable for different applications and scenarios, and they do not always tell the same story.

We then use our findings on measurement to conduct experiments addressing a broad question about how the use of transfer learning affects the fairness of a system. There is no previous work on this, but previous work on aspects of transfer learning leads to two competing possibilities of how transfer learning could impact fairness. Does transfer learning *improve* fairness, because the additional data sources lead to overall better models that are better at modelling long tail phenomena (and data on minorities is often long tail)? Or does the additional complexity bring in new or magnified undesirable biases, via one of the many mechanisms introduced above?

In Part II we pick a task—sentiment analysis, which we selected since this task enables us to test in a number of languages—and study this effect for transfer learning between *tasks/objectives* (the current dominant NLP paradigm, which we will sometimes refer to as monolingual transfer learning to distinguish it) and transfer learning between *languages*, called multilingual or crosslingual transfer learning (used interchangeably but the field and by us). Prior to our first investigation, previous work had shown that language models trained on unstructured text have gender and racial biases (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2019, 2020; Sheng et al., 2019). So we asked, will this carry through in monolingual transfer learning and cause gender and racial biases to appear or increase in a downstream sentiment model, beyond what can be attributed to the downstream training data? For instance, let's say that an upstream language model has learnt to associate conventionally negative attributes with certain minorities, such as to represent gay men as doing drugs, and black men as pimps (examples from Sheng et al. (2019)). Will a sentiment classifier built on this upstream language model also associate negative sentiment with gay and black men, *even if* there is little or no data about gay and black men in the sentiment training data? Or will that bias be overridden or lost, either because the role of the classifier is strong enough to disregard that, or because the now larger and more expressive system can generalise better to other positive association involving black and gay men, such as stars in politics and arts, or affirmational personal stories, such as those in Dixon et al. (2018)? We find that, overall, the additional stability from transfer learning is helpful in a resource constrained setting (i.e. one in which you cannot gather more annotated sentiment data), and this effect is enough to reduce overall gender and racial biases (despite new negative associations having been introduced).

We also study this effect for transfer learning between languages, or **cross-lingual transfer learning**. In this setting, not only can an upstream model learn biases from multiple data sources, but also from multiple languages. Exactly how much information cross-lingual transfer learning shares across languages is not well understood and there are some contradictory empirical studies (Conneau et al., 2020; Artetxe et al., 2020). We ask, in cross-lingual transfer learning, if a language model has learnt harmful stereotypes in one language, can those negative associations carry across languages? In the above example where a model has learnt negative associations *in English* about black and gay men, will a classifier in Japanese have these same associations, if they do not occur in Japanese? Can the collision between competing stereotypes in different languages weaken them, and in effect fight bias with bias? (Stanovsky et al., 2019). Can anything be done in the initial task before transfer, to ensure better outcomes in the second task? We find that, contrary to what we found in monolingual transfer learning, cross-lingual transfer learning tends to (with exceptions) exacerbate biases, though this effect can be mitigated with distilled/compressed models with little loss in performance.

In Part III, we look at a third type of system: retrieval augmented generation, which presents an inversion of the standard transfer learning setup. In the standard setup, a language model feeds into a classifier, and in retrieval augmented generation, the classifier selects source documents to answer a query, and this feeds into a language model, which conditions on those documents to generate an answer. This inverted system allows us to also ask the reverse question: if a language model has learnt problematic associations and stereotypes, can these be counteracted by conditioning on source documents? For instance, if a language model generates results about women predominantly in low-prestige roles, will it change this if it is conditioned on source documents about female CEOs and doctors? Or is it more likely to ignore the source information in this case then in the case of male CEOs and doctors? Or, as a third option, the retriever itself is biased, and doesn't select documents about female CEOs, so we never even get to that point?

However, prior to our work, not only was there no research examining how fairness flows between retrieval models and generative language models, there was little research analysing neural retrievers at all. So we began by asking the sub-question, inspired by all our work in Parts I and II: a retriever representation is necessarily a compression of a document, so what information is actually in this representation, such that a language model can condition on it? (Recall Chapter 4 where information in a representation as measured by information theoretic probing is most predictive of bias). Is information about demographics–gender, race, etc–in a retriever representation predictive of allocational bias in retrieved results? That is, does a retriever with stronger information about gender pick documents about gender more unequally? We do a case study in allocational gender bias and find that, though retrievers quite strongly encode gender in their representations, allocational bias is not attributable to the representations themselves. This bias persists even when we remove gender from the representation, meaning that it comes from either the composition of the corpus or the queries themselves.

We note that we define (un)fairness above generally, to highlight that our methods could apply to any subgroup divisions. In practice in our explorations, we look at either gender, race, immigrant status, or country of origin. We choose these types of bias as they either had pre-existing data, or lent themselves well to generating our own synthetic data, or labelling our own natural data. E.g. if we are given a few English paragraphs describing a person, it is usually easy to label their gender, compared to other features about them. This focus is not unique to our work: gender and race are the types of bias most represented in research for this reason (Goldfarb-Tarrant et al., 2023). But formally, we define (un)fairness broadly since everything we do can be applied to any axis of bias; to other protected characteristics like religion and orientation, or to non-protected characteristics that can cause big disparities in NLP systems, like regional or non-prestige dialects. We further note that our definition of unfairness is not NLP specific, such that it could be applied to other modalities as well: this is a larger extension of the work than to new bias types, but is possible since most of this research operates directly on representations. Each section of work will be extensible to further bias types, and potentially further modalities, and we encourage future work to do so.

1.1 Contributions

We make contributions to three broad categories:

1. More meaningful and reliable **measurement** of fairness in language models
2. Analysis of how **transfer learning** affects fairness
3. Analysis of fairness in **retrieval-augmented generation**

1.1.1 Measurement

Chapter 3: Intrinsic Bias Metrics Do Not Correlate with Application Bias

- We did the first study evaluating whether the most commonly used fairness metric for upstream language models correlated with downstream fairness. At the time, upstream only studies comprised one third of fairness research (Blodgett et al., 2020).
- We examined a much broader scope of experimental settings than most fairness research at the time. We looked at the relationship between upstream and downstream metrics across: two types of bias (gender, racial), two different tasks (coreference resolution and hatespeech detection), two different languages (English and Spanish), two common embedding algorithms (fastText and word2vec), two common methods of debiasing (preprocessing training data, and post-processing on representations), and two downstream fairness metrics (difference in precision and difference in recall).
- We found that the common upstream metric, based on cosine similarity, was **not** predictive of downstream bias. This changed the focus of the fairness field as a whole toward evaluating bias downstream, and towards finding alternative

upstream metrics that are more predictive. Our work has inspired follow up studies examining the predictive validity of fairness metrics (Cao et al., 2022), which further extend and corroborate our findings in other settings.

Chapter 4: How Gender Debiasing Affects Internal Model Representations, and Why It Matters

- We also did the first study investigating how debiasing *downstream* (rather than upstream) affects language model (upstream) representations.
- We focused on gender bias in English and considered two common transformer models, two tasks (coreference resolution, biography classification), three debiasing methods, two different intrinsic metrics: a contextual extension of the cosine similarity metric from the previous work and a new one, MDL compression, that we proposed adapted from Voita and Titov (2020). We looked at ten downstream fairness metrics, the largest number of which we are aware in a fairness study.
- We found that our new proposed metric was predictive of whether the upstream model had been successfully debiased, and correlated well with most downstream metrics.
- We also found that not all downstream fairness metrics correlated to each other, highlighting the importance of not relying overly much on one metric.

1.1.2 Transfer Learning

Chapter 5: Bias Beyond English: Counterfactual Tests for Bias in Sentiment Analysis in Four Languages
and
Chapter 6: Cross-lingual Transfer Can Worsen Bias in Sentiment Analysis

- We did the first research on the effect of both standard (monolingual) transfer learning and cross-lingual transfer learning on gender and racial biases in sentiment analysis.
- We first examined whether, for five languages (Japanese, Chinese, Spanish, German, English) monolingual transfer learning via pre-trained models changed the biases in sentiment analysis systems.

- We found monolingual transfer learning usually reduces biases, even though the training data used for transfer contains new biases. It stabilises the model and that effect outweighs bad content learnt in pre-training.
- We then ran similar experiments for the much more complex setup of multilingual transfer learning: via multilingual models and via cross-lingual labelled data.
- We found that, though the story is reasonably complex, cross-lingual transfer learning *can* increase bias even in unexpected cases such as culturally specific racial biases, which previously would've been expected to not transfer.

1.1.3 Retrievers

Chapter 7

- We did the first analysis of the properties of **Dense Retrievers** (as contrasted with sparse TF-IDF based approaches), which are the basic component of retrieval-augmented generation (RAG) systems. Knowing what information is in a retrieved representation is a pre-requisite to analysing how the retriever influences a downstream generative language model, but there was previously no work applying analysis or interpretability methods to retrievers.
- We analysed how the information captured in a representation differs for a retriever vs. the language model it was initialised from. We used information theoretic probing (based on the results in Chapter 4 that is was predictive of bias) to analyse how extractable two features were from a representation: topic of a passage and gender of a subject.
- We analysed how these features correlated to raw performance and to allocational gender bias. We found that gender extractability did correlate to performance on gender related questions and allocational gender bias, but that allocational gender bias persisted even when gender information was erased, meaning it was not attributable to the representation itself. We thus show another case when an entire system has to be considered in debiasing an NLP system.

1.2 Recommendations

In light of this body of research, we make the following recommendations.

On **Measurement**, we recommend not to use geometric intrinsic measurements of bias (based on cosine-similarity like WEAT (Caliskan et al., 2017) and CEAT (Guo and Caliskan, 2021)), as they are not predictive of downstream behaviour. This is true regardless of whether they are applied to a non-contextual embedding like word2vec (Mikolov et al., 2013a), or to a language model like BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) and company. These metrics *are* good for studying human social biases via what is reflected in the data that trained the model, as was done in the original work of Caliskan et al. (2017) that inspired the usage of this type of metric.[7] But they are not good for predicting *model* behaviour.

We can tentatively recommend instead using information theoretic probing as an alternative and reliably predictive intrinsic metric. However, this recommendation comes with two limitations: we studied information theoretic probing only for *allocational gender bias in English*. First, gender encoding differs greatly in different languages (more than other demographics) due to gender agreement systems, so these findings should be validated in more languages before being trusted beyond English. Second, even including English, other biases may not be stored the same way (for the same reason of the grammaticality of gender). So for other types of bias, no intrinsic metric has yet been validated and downstream metrics should still be used until more research has been done. Research on other options for intrinsic measurements is nascent, and we recommend always measuring fairness on a downstream task rather than in a language model when possible.

We also recommend that downstream metrics be selected with reference to the desired system behaviour. This may seem simple, but few works in the NLP literature acknowledge this, despite that the suite of all downstream fairness metrics is provably not mutually satisfiable, so you do actually have to pick one as a constraint. Different downstream metrics mean different things, and debiasing efforts often will only make sense for some metrics. Equalised false positive rates make more sense in the context of content moderation or toxicity, where the risk is censorship, equalised false negative rates make more sense for resume screening where the risk is excluding people from

[7] Though for this type of use case we note that RIPA (Ethayarajh et al., 2019) is likely better, or at the very least word frequencies need to be normalised for results to be valid.

the potential to interview.

In NLP, we often try to avoid making normative decisions about the world that our models will be embedded in; it is a messy and complex world, even more so than our data. Part of the brittleness and unreliability of bias evaluations and bias metrics—poor predictive and concurrent validity—is that researchers don't always think these through and make them explicit. Each debiasing method only make sense for some type of bias, and our better intrinsic metric from Chapter 4 still only correlates with most (not all) extrinsic measures; there is a family of measures that it does not work for. Fairness researchers do need to engage with the world they are imagining and how they believe it should function. All fairness work contains an assertion like this, and if left implicit, it can be scientifically messy. So we recommend that researchers make explicit, reasoned choices about the harms they are measuring and why they chose the metrics that they do.

On **Transfer learning**, we recommend to use monolingual transfer learning (also called pre-training) to augment lower-quantity supervised data, at least for classification tasks. We tested sentiment classification in three language families, so we expect our findings to hold for all similar tasks, but cannot claim to generalise to generative tasks.

However, we recommend to take more care when using cross-lingual transfer learning, as it risks introducing new biases into the target language from other language data. When cross-lingual transfer learning is used, we recommend using distilled cross-lingual models, as we found distilled models to have nearly equivalent performance and much lower bias overall than their full-size counterparts.

We recommend also the use of two of our analytical methods: causal or counterfactual evaluations, combined with a granular heatmap based analysis of the results.

On **Retrievers**, we recommend to analyse the entire system: corpus, queries, and model representations, as our work shows that a model constrained to have perfectly fair representation may still create an unfair system because of the other components. From the extensive experiments on random seed initialisations in this section, and the smaller scale experiments in the previous, we also recommend to test models based on a large number of random initialisations. We found this to have a disproportionate effect on model fairness and model performance both. In cases where trustworthy evaluations are available, ones which are faithful to a use case and which generalise, they can be used to select a seed with better generalisation properties for fairness, and this difference can exceed the difference from any common debiasing approaches or interventions. In cases where this is not possible, we recommend using majority voting across three to five random seeds, to minimise by seed variance.

Chapter 2

Background

The following sections give requisite background information common across either all works or multiple works in this thesis. Background that is relevant to only an individual work (probing methodologies, retrieval augmented generation, etc) will instead appear directly before that work.

2.1 Defining Fairness

Fairness is a relatively recent subject of study within the field of Machine Learning/AI research. As such it suffers from lack of standardisation in both definitions and methods of measurement. This is much to the detriment of this growing field. Many works fail to concretely define fairness (Blodgett et al., 2020; Goldfarb-Tarrant et al., 2023). Often when doing a meta-analysis or review of fairness literature, it is unclear if conflicting results are the result of a methodological problem, an error in code or analysis, or just a disagreement in definitions and the set of works actually should not be compared.To avoid these pitfalls, all work in this thesis will concretely define what fairness means in the context of that particular research. As background to all of them, I will give a brief overview of the discipline of fairness in NLP, how it has grown, and what 'fairness' tends to mean in different contexts. Fairness in AI began to gain attention in 2016, following the publication of a few high profile works within Machine Learning. The first was the popular book *Weapons of Math Destruction* (O'Neil, 2016), an exposé of all the ways that Machine Learning systems are invisibly incorporated into parts of our society, and how the assumptions baked into them propagate injustice. The second is the NeurIPS research paper from the same year, *Man is to Computer Programmer as*

Woman is to Homemaker? Debiasing Word Embeddings (Bolukbasi et al., 2016). This paper showed, via evocative word analogies from neural word embeddings (which was the standard measure of embedding performance at the time (Mikolov et al., 2013b; Finkelstein et al., 2002; Drozd et al., 2016)), how career-based gender bias was learnt by these systems, even when trained on relatively innocuous (for the internet) data like Google News. The field of Machine Learning was galvanised by these works, and more work began to be done on fairness analysis and mitigation within the following few years.[1] But definitions and methods are still being solidified.

2.2 Measuring Fairness

Notation: In all fairness metric definitions contained in this work, let $a \in A$ be the demographic variable in question, where $A = \{privileged, minoritised\}$ group, such as $\{male, female\}$ or $\{native, immigrant\}$[2] In classification tasks (all tasks until Part III) let Y be the true label, $\hat{Y}$ be the predicted label, and R be the classifier score (which enables analysis independent of classifier threshold).

Representational fairness has no codified metrics of measurement in NLP. This one of the clearest areas where NLP could learn from sociology and psychology, for they have been measuring representational fairness in media for quite some time (Dixon, 2017; Entman, 1992), but we've yet to operationalise this in NLP research. NLP largely neglected to measure representational fairness until Sheng et al. (2019), which proposed using a classifier to detect *regard* for the subject of a passage in open-domain generation. *Regard* captures how a reader of a text would esteem the subject. Using regard, they found GPT-2 (Radford et al., 2019) to systematically generate content causing lower regard when generating about women, African-Americans, and gays. This work is conceptually satisfying, and important, but difficult to expand due to the reliance on the classifier, which 1) is limited to English and 2) can become out of date over time as language drifts (and at the time of writing already has). So there have been not many follow up replications of this work but it doesn't have broad adoption (Goldfarb-Tarrant et al., 2023).

[1] It is worth noting at this stage that other fields had been aware of fairness issues in automated systems for some time, as education and hiring had been looking at statistical fairness for the previous half a century (Hutchinson and Mitchell, 2019). ML works built off of this, though perhaps not as much as they should've, and we did do some reinventing of the wheel.

[2] An obvious limitation of this is that privileged and minoritised is binary. This tends to be true of fairness work, including work in this thesis. There is insufficient work on extending fairness metrics and constraints to multiclass, either theoretically or empirically.

Approaches that differ from Sheng et al. (2019) tend to use sentiment score of the text (Goldfarb-Tarrant et al., 2023), and the current cutting edge Large Language Model (LLM) work still does Touvron et al. (2023); Jiang et al. (2024).[3] This is unfortunate, since the relationship of sentiment to representational harm is not well-correlated; many stereotypes that are harmful in a societal or an HR context can have positive sentiment (e.g. women are nurturing) (Fraser et al., 2021). The field is overdue for an analysis of the impact of this difference.

Other work on representational fairness that avoids using sentiment or regard classification focuses on discovery of language model stereotypes via challenge sets or customised prompts, and the likelihood of different generations (Smith et al., 2022). Challenge sets like this makes up the majority of bias work done on generative models today (Goldfarb-Tarrant et al., 2023), as generative fairness tends to focus on representational harms. However, the most prevalent approaches to stereotype measurement for the past few years, from two benchmark datasets, have been shown to be so flawed in construction as to be essentially meaningless (Blodgett et al., 2021). More recently, better and more reliable datasets and examinations have come out for representational fairness (Esiobu et al., 2023; Hosseini et al., 2023; Smith et al., 2022; Dhamala et al., 2021). But for the timeline of the work in this thesis, as a result of lack of consensus and good resources, this thesis focuses on only **allocational fairness**.

The most comprehensive overviews of strategies for measurement of **allocational fairness** are Hutchinson and Mitchell (2019) and Barocas et al. (2019). I will explain a subset of these that are important in this thesis. At a high level, allocational fairness can be measured as **individual fairness**, which answers the question ‘are the results for similar individuals equivalent’ and **group fairness**, or ‘is the performance for demographic subgroups equivalent’. In the former, the work lies in defining the similarity function. *What is similar? Are two individuals with the same university degree similar? Or only if you bucket by university prestige?* Individual similarity requires that you decide what does matter for similarity and what does not. In the latter, the work lies in selecting the demographic slices (what *are* the subgroups that should be equal?), and in choosing the performance measure. Choosing the demographic slices does not get much attention in NLP literature. There are a few nods to intersectionality (Subramanian et al., 2021; Ma et al., 2023; Lalor et al., 2022; Kearns et al., 2018) (i.e.

[3] This is itself a bit interesting, because the benchmark that current LLMs use, Dhamala et al. (2021), notes the limitations of most automated metrics, and uses a combination of sentiment, regard, and toxicity, as well as other two other metrics that they define. But this nuance seems to have been lost.

'your groups may be more complicated') and to unsupervised demographic group discovery (Zhao and Chang, 2020): otherwise works assume that demographic groups are given, gold standard, and that discrimination against different demographic axes is independent–i.e. discrimination against women can be treated entirely separately from against African Americans. This is patently false, gender and racial biases are interdependent and cause new, distinct bias effects when they intersect (Borenstein et al., 2023). There is much more attention given to how to measure performance disparity. Most NLP work uses group fairness, and measures this performance disparity. In classification, sometimes difference in F1 is used (Zhao et al., 2018) but many works use more granular measures such as **equalised odds** (Hardt et al., 2016) which enforces equal false-positive rates (FPR) and true-positive rates (TPR) across groups.

$$P(\hat{Y} = 1, A = a, Y = 0)(FPR) \tag{2.1}$$

$$P(\hat{Y} = 1, A = a, Y = 1)(TPR) \tag{2.2}$$

where 2.1 and 2.2 should be equal $\forall a \in A$.

Note that the second constraint 2.2 is equivalent to recall, as recall can be expressed the same way:

$$\frac{\hat{Y} = 1 | Y = 1}{(\hat{Y} = 1 | Y = 1) + \hat{Y} = 0 | Y = 1} \tag{2.3}$$

. The second constraint (recall) is often used in isolation as **equality of opportunity** (Hardt et al., 2016), a relaxation of **equalised odds**.

Occasionally some works include related but different group fairness metrics, discussed in Barocas et al. (2019), such as **independence**, **separation**, and (rarely) **sufficiency**.

In Chapter 3 we look at differences in recall and in precision, the former is equivalent to 2.3 above. In 4, we use a broad number of metrics: difference in True Positives, difference in False Positives, difference in Precision, difference in F1, Independence, Separation, Sufficiency. The first author of that work goes on in Orgad and Belinkov (2022) to show how there is often poor coverage of metrics that could be used for a given dataset or task, based partly on what we learnt in this work.

It is valuable for fairness analysis to report a broad set of metrics because fairness metrics in practice should be chosen based on the tasks in question. Choices of fairness metrics involve a normative judgment, whether implicit or explicit, though most research fails to acknowledge this. This is also too often left implicit, or made based on what some prior similar work has used, even if a different metric is both able to be used and would be more suited (Orgad and Belinkov, 2022). In other words, the logic of why one might choose one or another metric is hidden.

The distinction between **allocational representational** fairness, and our choice to focus on metrics for the former, corresponds almost directly to the type of NLP task. Allocational fairness is used for discriminative tasks and representational for generative tasks. This correspondence provides another insight into why representational fairness measurement is less developed than allocational: *all* evaluation for generative models is challenging and less reliable than for discriminative models, both automated evaluation (Novikova et al., 2017; Saphra et al., 2023) and human evaluation (Clark et al., 2021; Hosking et al., 2024). It is even more task dependent than model dependent; within one model architecture (such as BERT), research will use allocational fairness when that model is used for classification, and representational fairness when it is used for next-word prediction. This mapping of discriminative/generative to allocational/representational is how NLP fairness research has played out in practice, but is not entirely inherent. One could measure allocational fairness of generative models; for instance, instead of classifying resumes as recruiter callback yes/no, a model could generate summaries of resumes to be read by recruiters, who then make the callback decision themselves.[4] The difference in quality of summaries for different demographics would be allocational fairness. But this notion of allocational fairness requires us to be able to measure a delta in generation quality, and measuring quality of generations is an unresolved area of its own. This fact, combined with the nascency of using generative models for tasks that could be, or used to be, discriminative, means that in practice allocational vs. representational fairness tracks along with discriminative vs. generative. If the generative NLP hype continues long enough, this relationship may weaken, and much of the allocational work here could be expanded into generative NLP.

In Part II we shift to using **invariance under a counterfactual** in a downstream task. **Invariance under a counterfactual** describes the assertion that model predictions should be invariant to a perturbation: the example sentence *My sister had a wonderful day today* should be classified as positive sentiment, and this should not change if it is perturbed to be textitMy brother had a wonderful day today.

[4] This is now a genuine practice since the advent of LLMs in common usage.

This approach is very different in the data it requires and the hypotheses it can and cannot prove, from the metrics discussed thus far. I describe the distinction between these two types of measurement as **interventional** for these, and **observational** for the previous metrics. I borrow this terminology from the medical field because it gives correct intuitions. Observational studies can tell you that phenomena is occurring, but not *why*. Your model could be worse at recall of toxic content targeted at women, but it could be because female targeted toxic content is more diverse in your dataset, and thus more difficult to detect. Or maybe this isn't the case, but your classifier is poorly calibrated for this group compared to others. Interventional studies, by contrast, make a change (a perturbation) and observe the difference. Invariance under a counterfactual sets up a tasks where a perturbation *should not* change a prediction, and then measures change as a failure. This approach was popularised in ML as a test for robustness to noise in vision tasks (Zheng et al., 2016); an image correctly classified as a leopard should not change to being labelled as a butterfly when just a few pixels change, or when some noise is added that is imperceptible to a human observer. In NLP this is less easy to do, because it is harder to assert that labels should not change when working with the discrete space of language. But it works well when done carefully for fairness, where we can assert that changing the race, gender, or other demographic information of a name on a resume (from Emily Johnson to Lakisha Brown, as was done in the real life study of Bertrand and Mullainathan (2004)) should not change an output label in a resume processing system. So invariance tests require these carefully paired data points, so they cannot generally be done on the same data as observational studies.[5] The benefit of this type of study is that you don't have to be careful in slicing data–when you get a result, you know why. The downside is that the noising has to be done with care to only perturb where invariance *should be true*. For this reason invariance under a counterfactual data is often synthetic, and may not be representative of the true distribution of data. This is the weakness of this method.

In the first work on measurement (3), we focus on gaps in precision and recall, as previous work upon which we built our analysis used F1 (Zhao et al., 2018), and factoring them out gives both more granular analysis and also comparability to the equality of opportunity measure (Hardt et al., 2016). In the second, we use the full suite of possible metrics. In Part II we use don't use a subgroup metric, but instead

[5] Note that Winobias (Zhao et al., 2018), which we use in both works in Part I, could have been framed similarly as a counterfactual (though one where the label *should* flip, so not an invariance test. It was not framed this way, and instead was framed as subgroup fairness where the groups were 'pro' and 'anti' stereotypical. But it could have been.

use counterfactual examples that perturb one demographic variable, where we make an invariance assumption that values should not change under this perturbation, and the magnitude of the change is our metric. This method does not fit cleanly into individual or subgroup fairness, as it can be analysed on an individual example (which we do) but those examples have also been constructed to stand in for a demographic. E.g. in the counterfactual example: *I made her feel relieved* vs. *I made him feel relieved*, *her* and *him* are individual instances of bias, but also are stand-ins for the concept of gender. In Part III we measure retrieval rather than classification, so we use performance gap in the most common retrieval metric. We then construct a separate experiment for causality even though we have to use an observational metric, which brings the measurement approaches in Parts I (observational) and II (interventional) together.

To end the measurement section, the two works on measurement were motivation by the following observation. It seems potentially obvious to state, but the main desirable characteristic of a measurement of fairness is that it **a**) accurately measures the concept that it purports to measure and **b**) has a reliable relationship to real world fairness. When **a** and **b** are both true, the measurement has **construct validity** – a multi-faceted concept in the field of measurement modelling from the social sciences (Jacobs and Wallach, 2021), that attempts to define and make explicit the gaps between conceptualisation (e.g. my model should not discriminate based on race) and operationalisation (e.g. the performance gap between different racial groups, as identified by dialect identification).[6]. Much of the work in Part I was motivated by my observation that these types of validity had not been examined and were assumed to be true. We thus set out to test them.

2.3 Common Approaches to Debiasing

Fairness literature, as well as measuring bias, will often propose methods of **debiasing**. Debiasing methods proliferate, but most new methods do not get widespread adoption, since they fail to build trust. Debiasing methods tend to be proven in only quite constrained settings, on only one or two models, only in English, and on a limited number of tasks.[7]. This thesis therefore focuses on analysis, and does not propose any new

[6] Formally, **a** corresponds to **content validity** and **b** to **predictive validity**, as sub-concepts of **construct validity**

[7] I would like to here allocate appropriate blame to publication venues for requiring ‘novelty’ such that new works tend to propose new methods rather than verifying existing methods, leading to the situation at the commencement of this thesis where we had zillions of methods that no one used

methods. However, we will briefly survey existing methods that are used in analysis in Part I and III.

Debiasing approaches fall into high level categories of where they occur in the lifecycle of training an NLP model: pre-, mid (during), and post. **Preprocessing**[8] approaches involve a processing step that modifies data before training a model, to reduce signal that can cause bias. For example, if a system used for resume filtering ought to be debiased with regard to binary gender, the data can be processed such that there is an equal co-occurrence of gender signifiers (pronouns, names, other words that encode gender information) alongside words that indicate profession or career information.[9] Preprocessing can be done on unstructured webtext that will be used to learn embeddings or train language models or on labelled data that is used for supervised finetuning. These are usually known as **dataset balancing**, and differences in dataset balancing approaches stem from both the chosen method of editing data and the axes along which the data is balanced (gender/profession, race/toxicity, religion/sentiment, etc). The method falls into broadly two approaches (Schwartz and Stanovsky, 2022). If there is enough data that some can be removed without much performance penalty (more commonly true of unstructured text), it can be subsampled such that there is less but more balanced data (Wang et al., 2019). Other approaches oversample data such that some data is repeated (Chawla et al., 2002) in order to overweight those examples because that they occur more frequently in training data. This suffers from lack of diversity in the minority class, so other works opt to do a third option and instead create synthetic data for the minority class to remedy this (Dixon et al., 2018; Zhao et al., 2018). While the preprocessing approach is most reliable, and best understood, it is only available to practitioners who actually train models, which was always a small class; increasingly smaller as models scale. There is also very little work that tries to balance multiple axes at once which highlights the biggest limitation of dataset balancing. To achieve a minimally biased representation, you have to simultaneously balance across all genders, races, ages, etc, which is increasingly infeasible in any non-synthetic approach. Simultaneously balancing multiple axes is a perfectly reasonable desire in real life applications, and for many cases is a required feature. Very few regulations (or systems of integrity) have the goal of systems that are gender-fair but racist

[8] I use the term preprocessing rather than pre-training to distinguish from the now common terminology of pre-training/finetuning

[9] This is never easy to do fully, but can be quite successful in English with relatively coarse processing. It is not so easy in languages with much more gender marking, and this area is heavily under-researched. Gonen et al. (2019) looks into using morphological analysers for this.

and ageist. This is clearly important future work.

Debiasing can be done in **postprocessing** as well, generally on representations, though there is some preliminary work investigating utilising decoding parameters (Sheng et al., 2021). Both approaches are more complex than preprocessing, both conceptually and in implementation, but do not require retraining a large and expensive model. Crucially, this enables debiasing to be done by parties further downstream who are then most connected to a downstream application.[10] It further allows more iteration and experimentation without extensive compute. Ravfogel et al. (2020) operate on representations via nullspace projection – they learn a linear classifier for a demographic (binary gender, race) and then project language model representations onto the nullspace of that classifier. Iskander et al. (2023) extend this method to remove non-linearly encoded information. We use these methods for causal analysis of the impact of demographics (rather than debiasing) in Part III. The other methods of debiasing representations operate on individual words and groups of words that stand in for concepts: Mrkšić et al. (2017) pushes word embeddings together or away from each other in representation space; we use this method in Chapter 3. More recently a number of works use model-editing (Meng et al., 2022), where individual neuron values can be changed in order to change one specific output string. This has some issues with scaling to a full demographic (edits are granular) but would be a promising new direction for very targeted interventions. It is a promising new direction, but post-dates the work in this thesis, and thus is not used.

Debiasing during training, via constraints or costs to the learning method (Zhao et al., 2017), is less commonly done, perhaps partly perhaps because it contains the disadvantages of both preprocessing and postprocessing – it is conceptually more complex and requires tuning hyperparameters (as does postprocessing) but it also requires retraining a model. It also is made more complex by Transfer Learning, as it is unclear whether to do it at one or both stages. The formalisation is satisfying, because explicit constraints give quantifiable fairness outcomes, but with increasing scale it is increasingly impractical. We do no analysis on this kind of debiasing for this reason.

Recent hype around generative language model fairness focuses on a second stage training process, often called ‘alignment’, which refers to the idea that human morality (it is never specified which human or which morality) can be instilled in a model via fine-tuning with a ranking loss over examples that are more or less moral. The majority

[10] We show this connection to downstream is necessary in Chapter 3.

of this thesis predates the alignment trend, and very little deals explicitly with generation, so we do not use any of these techniques. We note also that much of the alignment work has origins in robotics more than in fairness. We do, however, discuss current implications of this work in the Conclusion (8). However, our work does have interesting similarities to point out at this stage. In Chapter 4 we measure fairness via differences in distributions for different demographics, via KL or Wasserstein distance, one of the standard ways to measure it (§2.2). This measurement is even sometimes directly optimised for (in a small violation of Goodhart's law) in works like Huang et al. (2020), who regularise output to have similar sentiment distributions between groups. Korbak et al. (2022) show that Reinforcement Learning from Human Feedback (RLHF), the most common method of alignment, can be equivalent to distribution matching and Rafailov et al. (2023) transition the implementation of this to a new objective that does this explicitly, and many use it for its superior stability and ease of implementation. So when given an appropriate setup, RLHF could (theoretically) be directly optimising for a fairness constraint. Again, this postdates this thesis, and we do no alignment at all, but found it worthwhile to highlight the theoretical continuity in the approaches.

2.4 Fairness as Dataset Artifacts or as Failure to Generalise

Fairness issues can often be seen as a special case of one of these two areas, though this is rarely discussed in most fairness work. As an example, take racial bias from AAVE, in the two forms that we have already discussed in the background and introduction. In one of the cases, racial bias in toxicity detection has come from the model learning a dataset artifact, where labelled training data correlated African American dialect (AAVE) features with toxic content, as a result of an error or a bias in annotation (Sap et al., 2019). But allocational racial bias can also come from insufficient training data in AAVE , resulting in higher error rates from that group, as Tatman and Kasten (2017) measure for automatic captioning. Most work does not address this difference or disentangle these two causes, and lump both under "bias".

This lack of distinction is one of the reasons fairness work can fail to have predictive validity. Correcting an anti-AAVE stereotype may not help allocational bias in a model if the root cause was simply that it modelled AAVE poorly. Even dataset artifacts can be further disentangled and split into two types of causes conceptually: dataset artifacts (or dataset biases) that replicate historical biases (most previous engineers hired were men, and so the dataset of successful resumes is mostly male resumes) and more indirect dataset artifacts such as a correlation between line-length of resumes and the 'hire' label, combined with a notable difference in resume length between different genders such that male resumes are more hireable, but only because of an odd artifact of this dataset that it is correlated with. In some sense these two are the same, they are detectable via similar methods, and are a shortcut to a real task caused by a particular dataset construction – but I make the distinction as I've found they may be differently anticipated by humans. One is predictable given knowledge of historic dataset biases, the other is difficult to anticipate, and often so surprising as to be comical, as when NLI contradiction could be largely predicted by the presence of words about cats (Gururangan et al., 2018). They belong in the same category as far as causal effects, but the conceptual difference can have an impact on discoverability.

The collapse of the two causes–dataset artifacts and failure to generalise–into one measure is not necessarily bad, since the behaviour in an application is the same, and thus the real world impact on people is the same. But it would be beneficial for researchers to develop ways to split out these two causes to better suggest mitigations. Splitting them out has another important benefit–it shows the overlap between fairness research and other areas of NLP. Dataset bias work has significant overlap with work on dataset artifacts and on 'shortcutting' (Geirhos et al., 2020), generalisation failures have overlap with research on robustness and generalisation (Hupkes et al., 2023). If we explicitly recognise and leverage this, the field can share approaches and progress quickly, more than is currently done. For instance, AFLite is an algorithm developed to search a dataset for artifacts (such as 'cat' and 'contradiction') (Bras et al., 2020) and then filter them, and comes from the dataset artifacts literature. LOGAN (Zhao and Chang, 2020) is an algorithm for unsupervised discovery of social biases, from the fairness literature. They are implemented differently, AFLite is conceptually similar to k-fold validation with targeted sampling for artifacts, and LOGAN is a modification of k-means. But they can both be used to solve the same goal of finding slices of a dataset that exhibit strong imbalances based on a feature that should not have an imbalance. Since this comes from different subfields, no one has compared them. Similarly, the aforementioned work on fairness showing that automatic captioning doesn't generalise well to accents beyond white Californian male accents (Tatman and Kasten, 2017) and that facial recognition doesn't generalise to non-white skin (Buolamwini and Gebru,

2018) has much conceptually in common with generalisation work showing that natural language inference doesn't generalise to new syntactic structures (McCoy et al., 2020). The areas do not acknowledge each other currently, nor share mitigation or analysis techniques, but there is much room to do so.

The field would benefit from the disentanglement of these two causes, but also from further disentanglement of other contributing factors. Both 'dataset artifacts' and 'failure to generalise' have two contributing factors: the data, and the model. As a result of the latter, they overlap with research on inductive biases, which inspires the work in Part III. In Chapter 1, I referenced the common misconception that application bias results solely from biases in training data. But reliance on dataset artifacts is not just from data skew, and failure to generalise is not just from insufficient data. Both can come from either the model or the data.

The exact same data with the same artifacts can result in a model that is strongly reliant on those artifacts, or a model that has not learnt them, and instead relies on other features. We found this in Chapters 4 and 7, and it was also found in Lovering et al. (2021) and Sellam et al. (2022). A perfect model of data with artifacts strongly correlated to labels would be biased, but an imperfect model of the same data could be more biased, with more skew (usually termed **bias amplification**) or could be less biased, if the artifact was not modelled well. Both of these worsening and lessening effects bias effects are shown in the spread of WinoBias scores in Sellam et al. (2022), resulting from different model inductive biases.

Even with perfect data (the admittedly less common possible case), an imperfect model could learn a function that will induce a skew and cause failure to generalise from the model alone. In this case, there could be plenty of non-white faces in a facial recognition training dataset but the model still may perform worse on them. Since it is not a data issue, adding more data and dataset balancing would not be effective mitigations. This illustrates how work in the fairness field is entangled both with work on data quality but also with work on inductive biases.

Given this observation, in this thesis I attempt to take inspiration and techniques from these related fields and incorporate them into fairness research. In Chapter 4 and Parts II and III we run all experiments on multiple random seed initialisations, and analyse the models separately by seed. This is rarely done in fairness work, but generalisation work has shown it to drastically affect results (McCoy et al., 2020; Sellam et al., 2022). Our results corroborate this; different seeds do show drastically different fairness properties despite equivalent development set performance, just as was found in McCoy et al. (2020).

We hope that in future these fields have more dialogue and joint work.

2.5 Transfer Learning

This thesis is on Fairness for Transfer Learning, so Transfer Learning also deserves some background explanation. Before this thesis, Transfer Learning had no intersection with fairness research. This has changed, partly because of the work here and follow on work it inspired, and partly because, Transfer Learning is, at the point of this thesis being written, so common that it is not generally specified anymore and is the unstated default. Back when this thesis was in its infancy, both the fields of fairness and of Transfer Learning were very small but growing exponentially. It seems that Transfer Learning has won, as almost everything is Transfer Learning (though fairness is no small field anymore either).

The central premise of Transfer Learning is that it doesn't make sense to start from a tabula rasa randomly initialised weight matrix every time you want to learn a new task, which we used to do before, but that many of the concepts necessary for one NLP task may be in common with another. Toxicity detection and sentiment analysis both require knowledge of basic sentence structure, nouns, verbs, and negative connotations of different words, and so knowledge from one should be able to augment knowledge from the other. Even more dissimilar tasks like toxicity detection and coreference resolution still require a similar underlying knowledge of sentence structure. Early work in Transfer Learning often sought to transfer from task to task like this; this approach is called *domain adaptation* when done in sequence, or *multi-task learning* when done simultaneously (Ruder et al., 2019). Both these approaches are now less common, and the field of Transfer Learning has coalesced into one paradigm: sequential pre-training of a language model on unlabelled text, followed by task specific fine-tuning. This leverages the vast amount of unstructured text to build high quality representations of language that can then serve as initialisations for any downstream language task. This is the approach that we use throughout this work.

For additional detail, in Part II, we use a variant of Transfer Learning that combines aspects of both the pre-train-finetune paradigm and the domain adaptation paradigm, when we do cross-lingual transfer. In cross-lingual Transfer Learning, the language model pre-training stage is multilingual (contains text in a variety of languages) and the fine-tuning stage is in a high-resource language (usually English) that doesn't match the target language at inference time. This is strictly called *zero-shot cross-lingual transfer* (ZS-XLT) since the model has never seen labelled data in the target language. There also exists *few-shot cross-lingual transfer* (FS-XLT), where the high-resource fine-tuning is continued with a few examples in the target language (often only hundreds). FS-XLT has been shown to generally perform better than ZS-XLT for relatively low additional annotation cost (Lauscher et al., 2020), but we use exclusively ZS-XLT in this work, since FS-XLT adds many additional layers of tuning and variability to the already complex landscape of cross-lingual transfer, and the small additional bump in performance was not necessary for our analysis.

Early Transfer Learning research varied between whether to 'freeze' the language model after the first stage, and just learn whatever new parameters need to be added for the desired final task output space (e.g. the classifier or coreference model or etc that takes in the representations) or to continue to train the language model along with the second task, to further refine the representations to best fit the task specific needs. In this work we use both methods, whichever is most standard for the task and enabled ease of analysis. We always specify which we use in each work's respective methodology.

Part I

Measuring the Relationship between Fairness in Pretraining and Fairness in Downstream Applications

In the sequential transfer learning paradigm – the dominant approach to transfer learning – a new difficulty for fairness research emerges. A language model, which is trained first, can be used in many different downstream applications. Any biases learnt by the language model, can propagate into many different applications. But testing in every single application would be onerous and is sometimes not possible: the engineer training a language model may not have the data and expertise to train models in toxicity detection and sentiment analysis and named entity recognition and information retrieval...and the many other tasks that are instantiated from a BERT model, still the most popular pretrained model at time of writing. Even if they can, they're unlikely to have the domain knowledge to reason about the types of algorithmic discrimination that are risks for each use case. Accurate testing on a full set of downstream applications is impossible. So it is desirable to have a way to measure bias in the first stage, at the language model level, and to be able to predict effects downstream.

The NLP field did quickly search for a way to do this. At the same NeurIPS in 2016 one of the keynotes was on the exponential growth of transfer learning (Ng, 2016), and Bolukbasi et al. (2016) published the first work on gender debiasing word embeddings using a post-processing method based on PCA. This work analysed bias via the lens of the word analogy task (man is to woman as king is to _), which had been quite popular as an assessment of word embedding performance. This was the start of considering bias to be a property of word embedding geometry. This was reinforced shortly thereafter, when a computational social science work showed that the Implicit Association Test (IAT) for human psychological biases could replicated via cosine similarities between word embeddings Caliskan et al. (2017). This new measure, WEAT, became used as a predictive measure of language model biases, and was used as the sole metric to support a plethora of new embedding debiasing algorithms.[11] In both of these, bias was operationalised as distances between word vectors in three hundred to one thousand dimensional space, or via finding the principal components associated with a demographic (usually gender) subspace (Ethayarajh et al., 2019).

The following work challenges the implicit assumption behind this measure – that a measurement of embedding geometry, WEAT, is predictive of downstream bias measurements. WEAT is observational, like the downstream extrinsic fairness metrics

[11] The method of Caliskan et al. (2017) was used in 213 different works at the time that our work came out, as counted by Semantic Scholar *Methods* citations, https://www.semanticscholar.org/paper/Semantics-derived-automatically-from-language-Caliskan-Bryson/5966d7c7f60898d610812e24c64d4d57855ad86a?year%5B0%5D=2017&year%5B1%5D=2021&sort=relevance&citationIntent=methodology.

discussed in §2.2, but unlike those metrics, it is not *directly* observational of societal harm, the way True Positive Rate gaps are. It is observational of an embedding space, but the relationship between that space and societal harm has not been established.

Prior to the work below it had only been established to reveal bias and imbalances in concepts in a dataset, as an additional tool for understanding sampling bias and reporting bias and even artifacts that are nonetheless in unstructured data and can be learnt from the collocational way that embeddings are trained. However, the body of bias work on embeddings used this measure with the assumption that it had predictive validity. Research implicitly assumed that downstream bias would track with WEAT metrics, or at the very least, that if WEAT bias measures went down (less bias) then downstream bias would go down.

Given the many possible definitions of bias and different underlying causes outlined in §2.2 and §2.4, this is quite a strong claim, too strong to assume to be true from intuition. On close examination in fact, intuition would not necessarily support this. Some tasks do rely on concept proximity, like recommendation engines at the time, and for these intuition would suggest that a measure of embedding geometry might be predictive of downstream task bias. But this is an exception. For most tasks that involve transfer learning, the objective function used for training and the output space for a downstream classifier are very different. This is one of the hallmarks of transfer learning and makes it complex. In this case it seems unlikely that a simple measure like WEAT will be consistently predictive of the variety of downstream tasks. The claim that debiasing an embedding space is helpful not only has unproven predictive validity, but lacks *face validity*, which is the ‘sniff test’ of whether it looks on the surface level to be plausible.

This observation motivates the following work: to discover if and when WEAT has predictive validity of bias in downstream applications, and the utility of using it as a measure.

Chapter 3

Intrinsic Bias Metrics Do Not Correlate with Application Bias

Intrinsic Bias Metrics Do Not Correlate with Application Bias

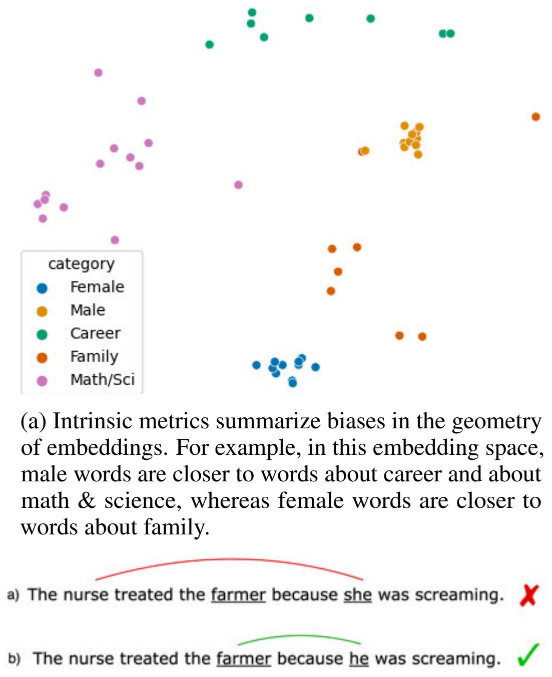
Seraphina Goldfarb-Tarrant[*†] **Rebecca Marchant**[*†] **Ricardo Muñoz Sánchez**[*†]
Mugdha Pandya[*†] **Adam Lopez**[‡†]
[†]University of Edinburgh, [‡]Rasa Technologies GmbH

Abstract

Natural Language Processing (NLP) systems learn harmful societal biases that cause them to amplify inequality as they are deployed in more and more situations. To guide efforts at debiasing these systems, the NLP community relies on a variety of metrics that quantify bias in models. Some of these metrics are *intrinsic*, measuring bias in word embedding spaces, and some are *extrinsic*, measuring bias in downstream tasks that the word embeddings enable. Do these intrinsic and extrinsic metrics correlate with each other? We compare intrinsic and extrinsic metrics across hundreds of trained models covering different tasks and experimental conditions. Our results show *no reliable correlation* between these metrics that holds in all scenarios across tasks and languages. We urge researchers working on debiasing to focus on extrinsic measures of bias, and to make using these measures more feasible via creation of new challenge sets and annotated test data. To aid this effort, we release code, a new intrinsic metric, and an annotated test set focused on gender bias in hate speech.[1]

1 Introduction

Awareness of bias in Natural Language Processing (NLP) systems has rapidly increased as more and more systems are discovered to perpetuate societal unfairness at massive scales. This awareness has prompted a surge of research into measuring and mitigating bias, but this research suffers from lack of consistent metrics that discover and measure bias. Instead, work on bias is "rife with unstated assumptions" (Blodgett et al., 2020) and relies on metrics that are easy to measure rather than metrics that meaningfully detect bias in applications.



(a) Intrinsic metrics summarize biases in the geometry of embeddings. For example, in this embedding space, male words are closer to words about career and about math & science, whereas female words are closer to words about family.

a) The nurse treated the farmer because she was screaming. ✗

b) The nurse treated the farmer because he was screaming. ✓

(b) Extrinsic bias metrics summarize disparities in application performance across populations, such as rates of false negatives between different gender groups. For example, a coreference system may make more errors in an anti-stereotypical career coreferent (red arc) than in a pro-stereotypical one (green arc).

Figure 1: The relationship between *intrinsic* bias metrics (a) and *extrinsic* bias metrics (b) has been assumed, but not confirmed.

A recent comprehensive survey of bias in NLP (Blodgett et al., 2020) found that one third of all research papers focused on bias in word embeddings. This makes embeddings the most common topic in studies of bias — over twice as common as any other topic related to bias in NLP. As is visualised in Figure 1a, bias in embedding spaces is measured with *intrinsic* metrics, most commonly with the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), which relates bias to the geometry of the embedding space. Once embeddings are incorporated into an application, bias can be measured via *extrinsic* metrics (Figure 1b) that

[*] Equal contribution. Correspondence to

[1] https://tinyurl.com/serif-embed

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing, pages 1926–1940
August 1–6, 2021. ©2021 Association for Computational Linguistics

test whether the application performs differently on language related to different populations. Hence, research on debiasing embeddings relies crucially on a hypothesis that doing so will remove or reduce bias in downstream applications. However, we are aware of *no* prior research that confirms this hypothesis.

This untested assumption leaves NLP bias research in a precarious position. Research into the *semantics* of word embeddings has already shown that intrinsic metrics (e.g. using analogies and semantic similarity, as in Hill et al., 2015) do not correlate well with extrinsic metrics (Faruqui et al., 2016). Research into the bias of word embeddings lacks the same type of systematic study, and thus as a field we are exposed to three large risks: 1) making misleading claims about the fairness of our systems, 2) concentrating our efforts on the wrong problem, and most importantly, 3) feeling a false sense of security that we are making more progress on the problem than we are. Our bias research can be rigorous and innovative, but unless we understand the limitations of metrics we use to evaluate it, it might have no impact.

In this paper, we ask: **Does the commonly used intrinsic metric for embeddings (WEAT) correlate with extrinsic metrics of application bias?** To answer this question, we analyse the relationship between intrinsic and extrinsic bias. Our study considers two languages (English and Spanish), two common embedding algorithms (word2vec and fastText) and two downstream tasks (coreference resolution and hatespeech detection).

While we find a moderately high correlation between these metrics in a handful of conditions, we find no correlation or even negative correlation in most conditions. Therefore, we recommend that the ethical scientist or engineer does not rely on intrinsic metrics when attempting to mitigate bias, but instead focuses on the harms of specific applications and test for bias directly.

As additional contributions to these findings, we release new WEAT metrics for Spanish, and a new gender-annotated test set for hatespeech detection for English, both of which we created in the course of this research.

2 Bias Metrics

In all of our experiments, we compute correlations between commonly-used metrics, both intrinsic and extrinsic.

2.1 Intrinsic bias metrics

Intrinsic bias metrics are applied directly to word embeddings, formulating bias in terms of geometric relationships between *concepts* such as *male*, *female*, *career*, or *family*. Each concept is in turn represented by curated wordlists. For example, the concept *male* is represented by words like *brother, father, grandfather,* etc. while the concept *math & science* is represented by words like *programmer, engineer,* etc.

The most commonly used metric is WEAT (Caliskan et al., 2017).[2], which measures the difference in mean cosine similarity between two *target* concepts X and Y; and two *attribute* concepts A and B. This difference represents the imbalance in associations between concepts. Using $\vec{w}$ to represent the embedding of word w, we have a *test statistic*:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where

$$s(w, A, B) = \underset{a \in A}{\text{mean}} \cos(\vec{w}, \vec{a}) - \underset{b \in B}{\text{mean}} \cos(\vec{w}, \vec{b})$$

This is normalised by the standard deviation to get the *effect size* which we use in our experiments.

WEAT was initially developed as an *indicator* of bias, to show that the Implicit Association Test (IAT) from the field of psychology (Greenwald et al., 1998) can be replicated via word embeddings measurements. There are thus 10 original tests chosen to replicate the tests presented to human subjects in IAT. The tests measure different kinds of biased associations, such as African-American names vs. White names with pleasant vs. unpleasant terms, and female terms vs. male terms with career vs. family words.

WEAT was later repurposed as a *predictor* of bias in embedding spaces, via a somewhat muddy logical journey. It has since been translated into 6 other languages (XWEAT; Lauscher and Glavas, 2019), and extended to operate on full sentences (May et al., 2019) and on contextual language models (Kurita et al., 2019). When WEAT is used as a metric, papers report the effect size of the subset of tests relevant to the task at hand, each separately.

There are known issues with WEAT, such as sensitivity to corpus word frequency, and sensitivity

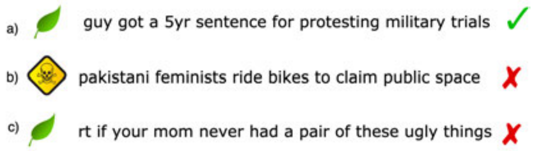
[2] We count 34 papers from *CL and FAT* conferences since January 2020 that use WEAT or SEAT (May et al., 2019) in their methodology.

to target and attribute wordlists, as found by Sedoc and Ungar (2019) and Ethayarajh et al. (2019). The latter proposes an alternative more theoretically robust metric, relational inner product association (RIPA), which uses the principal component of a gender subspace (determined via the method of Bolukbasi et al. (2016)) to directly measure how ”gendered” a word is. We have chosen to use the most common version of WEAT for this first empirical study, since it is most widely used. It would be interesting to test RIPA in the same way, if it were extended to more types of bias and more languages. But we note that all intrinsic metrics are sensitive to chosen wordlists, so this must be done carefully, especially across languages, a topic we will return to in Section 4.3.

2.2 Extrinsic bias metrics

Extrinsic bias metrics measure bias in applications, via some variant of performance disparity, or *performance gap* between groups. For instance, a speech recognition system is unfair if it has higher error rates for African-American dialects (Tatman, 2017), meaning that systems perform less well for those speakers. A hiring classification system is unfair if it has more false negatives for women than for men, meaning that more qualified women are accidentally rejected than are qualified men.[3] There are two commonly used metrics to quantify this possible performance disparity: Predictive Parity (Hutchinson and Mitchell, 2019), which measures the difference in *precision* for a privileged and non-privileged group, and Equality of Opportunity (Hardt et al., 2016), which measures the difference in *recall* between those groups (see Appendix A for formal definitions).

The metric that best identifies bias in a system varies based on the task. For different applications, false negatives may be more harmful, for others false positives may be. For our first task of coreference (Figure 1b), false negatives — where the system fails to identify anti-stereotypical coreference chains (e.g. women as farmers or as CEOs) — are more harmful to the underprivileged class than false positives. For our second task, hate speech detection (Figure 2), both can be harmful, for different reasons. False positives for one group can systematically censor certain content, as has been found for hate speech detection applied to African-American Vernacular English (AAVE) (Sap et al., 2019; Davidson et al., 2019). False negatives permit abuse of minority populations that are targets of hate speech. We examine performance gaps in both precision and in recall for broad coverage.



a) guy got a 5yr sentence for protesting military trials ✓

b) pakistani feminists ride bikes to claim public space ✗

c) rt if your mom never had a pair of these ugly things ✗

Figure 2: Examples from twitter hatespeech detection: correct (a), false positive (b), and false negative (c). This shows both kinds of problematic performance gap. b) censors harmless text and c) lets a targeted toxic comment slip through.

3 Methodology

Each of our experiments measures the correlation between a specific instance of WEAT and a specific extrinsic bias metric. In each experiment, we train an embedding, measure the bias according to WEAT, and measure the bias in the downstream task that uses that embedding. We then modify the embeddings by applying an algorithm to either *debias* them, or — by inverting the algorithm's behavior — to *overbias* them. Again we measure WEAT on the modified embedding and also the downstream bias in the target task. When we have done this multiple times until we reach a stopping condition (detailed below), we compute the correlation between the two metrics (via Pearson correlation and analysis with scatterplots).

Rather than draw conclusions from a single experiment, we attempt to draw more robust conclusions by running many experiments, which vary along several dimensions. We consider two common embedding algorithms, two tasks, and two languages. A full table of experiment conditions can be found in Table 1.

3.1 Debiasing and Overbiasing

We need to measure the relationship between intrinsic and extrinsic metrics as bias changes, we must generate many datapoints for each experiment. Previous work on bias in embeddings studies methods to *reduce* embedding bias. To generate enough data points, we take the novel approach of both decreasing *and* increasing bias in the embeddings. We measure the baseline bias level, via WEAT, for each embedding trained normally on the original corpus. We then adjust the bias up or down, remeasure WEAT, and measure the change in the downstream task.

[3] https://tinyurl.com/y6c6clzu

Task	Data	Bias Type	Intrinsic Metrics
English Coreference	Ontonotes/WinoBias	Gender	WEAT 6, 7, 8
English Hate speech	Twitter	Gender	WEAT 6, 7, 8
Spanish Hate speech	Twitter	Gender	XWEAT 7+8 (new)
Spanish Hate speech	Twitter	Migrants	XWEAT Migrants (new)

Table 1: Tasks used in our experiments. Each experiment consists of a task, an embedding method (either word2vec or fasttext), an intrinsic metric (one experiment for each listed), and an extrinsic metric (either Predictive Parity or Equality of Opportunity). We run an experiment for all possible combinations. To produce data points for each experiment, we use preprocessing and post-processing methods to debias and overbias the input word embeddings.

We choose two methods from previous work that are capable of both debiasing and overbiasing: the first is a preprocessing method that operates on the training data before training, the second is a post-processing method that operates on the embedding space once it has been trained. This is important since both kinds of methods may be used in practice: a large company with proprietary data will train embeddings from scratch, and thus may use a preprocessing method; whereas a small company may rely on publicly available pretrained embeddings, and thus use a post-processing method. [4]

For preprocessing, we use dataset balancing (Dixon et al., 2018), which consists of sub-sampling the training data to be more equal with respect to some attributes. For instance, if we are adjusting gender bias, we identify pro-stereotypical sentences[5] such as '*She* was a talented housekeeper' vs. anti-stereotypical sentences, such as '*He* was a talented housekeeper' or 'She was a talented analyst'. We sub-sample and reduce the frequency of the pro-stereotypical collocations to debias, and sub-sample the anti-stereotypical conditions to overbias.

As a postprocessing method for already trained embeddings, we use the Attract-Repel (Mrksic et al., 2017) algorithm. This algorithm was developed to use dictionary wordlists (synonyms, antonyms) to refine semantic spaces. It aims to move similar words (synonyms) close to each other and dissimilar words (antonyms) farther from each other, while keeping a regularisation term to preserve original semantics as much as possible. Lauscher et al. (2020) used an approach inspired by Attract-Repel for debiasing, though with constraints implemented somewhat differently. We use the same pro- and anti-stereotypical wordlists as in dataset-balancing. For debiasing, we use the algorithm to increase distance between pro-stereotypical combinations (*she, housekeeper*) and decrease distance between anti-stereotypical combinations (*she, analyst* or *he, housekeeper*). For overbiasing we do the reverse.[6]

As the stopping condition for preprocessing, we constrain the sub-sampling so that it does not substantially change the dataset size, by limiting it to removing less than five percent of the original data. For postprocessing we limit the algorithm to maximum 5 iterations.

3.2 Embedding Algorithms

We use two common word embedding algorithms: fastText (Bojanowski et al., 2017) and Skip-gram word2vec (Mikolov et al., 2013) embeddings. Word embeddings in fastText are composed from embeddings of both the word and its subwords in the form of character n-grams. Lauscher and Glavas (2019) suggest that this difference may cause bias to be acquired and encoded differently in fastText and word2vec (We discuss this in more detail in Section 5).

Despite recent widespread interest in contextual embeddings (e.g. BERT; Devlin et al., 2019), our experiments use these simpler contextless embed-

[4] There are additional embedding based debiasing methods used in practice, based on identifying and removing a gender subspace during training or as postprocessing (Bolukbasi et al., 2016; Zhao et al., 2018b). However, these methods do not change a word's nearest neighbour clusters (Gonen and Goldberg, 2019), and so we would expect these debiasing methods to show superficial bias changes in WEAT without changing downstream bias. Both methods that we select modify the underlying word distribution and move many words in relation to each other. We verified this with tSNE visualisation as in Figure 1a following Gonen and Goldberg (2019) and find that our bias modification methods do change word clusters.

[5] Stereotypes as defined by Zhao et al. (2018a) and by Caliskan et al. (2017), who use the U.S. Bureau of Labor Statistics and the Implicit Association Test, respectively.

[6] Wordlists used for bias-modification and configs for Attract-Repel are included in the codebase.

dings because they are widely available in many toolkits and used in many real-world applications. Their design simplifies our experiments, whereas contextual embeddings would add significant complexity. However, we know that bias is still a problem for large contextual embeddings (Zhao et al., 2019, 2020; Gehman et al., 2020; Sheng et al., 2019), so our work remains important. If intrinsic and extrinsic measures do not correlate with simple embeddings, this result is unlikely to be changed by adding more architectural layers and configurable hyperparameters.

3.3 Downstream tasks

We use three tasks that appear often in bias literature: Coreference resolution for English, hate speech detection for English, and hate speech detection for Spanish. To make the scenarios as realistic as possible, we use a common, easy-to-implement and high performing architecture for each task: the end-to-end coreference system of Lee et al. (2017) and the the CNN of Kim (2014), which has been used in high-scoring systems in recent hate speech detection shared tasks (Basile et al., 2019). For each task, we feed pretrained embeddings to the model, frozen, and then train the model using the standard hyperparameters published for each model and task.

3.4 Languages

We experiment on both English and Spanish. It is important to take a language with pervasive gender-marking (Spanish) into account, as previous work has shown that grammatical gender-marking has a strong effect on gender bias in embeddings (McCurdy and Serbetci, 2017; Gonen et al., 2019; Zhou et al., 2019). We use Spanish only for hate speech detection, because gender marking makes a challenge-set style coreference evaluation trivial to resolve and not a candidate for detection of gender bias.[7]

4 Experiments

4.1 Datasets

To train embeddings, we use domain-matched data for each downstream task. For coreference we train on Wikipedia data, and for hatespeech detection we train on English tweets or Spanish tweets, consistent with the task.[8] Our English Coreference system is trained on OntoNotes (Weischedel et al., 2017) and evaluated on Winobias (Zhao et al., 2018a), a Winograd-schema style challenge set designed to measure gender bias in coreference resolution. English hate speech detection uses the abusive tweets dataset of Founta et al. (2018), and is evaluated on the test set of ten thousand tweets, which we have hand labelled as targeted *male*, *female*, and *neutral* (we release this labelled test set for future work). Spanish hate speech detection uses the data from the shared task of Basile et al. (2019), which contains labels for comments directed at women and directed at migrants.

4.2 WEAT & Bias modification wordlists

Both WEAT and bias modification methods depend on seed wordlists.[9] These wordlists are closely related to each other, and we match them by type of bias, such that we measure WEAT tests for gender bias with embeddings modified via gender bias wordlists (themselves derived from WEAT lists, as detailed below) and WEAT tests for migrant bias with embeddings modified for migrant bias.

WEAT wordlists are standardised, and for English we use the three WEAT test wordlists (numbers 6,7,8) for gender.[10]

To generate bias modification wordlists we follow the approach of Lauscher et al. (2020) and use a pretrained set of embeddings (from `spacy.io`) to expand the set of WEAT words to their 100 unique nearest neighbours. For all experiments, we take the union of all WEAT terms, expand them, and use this expanded set for both dataset balancing and for Attract-Repel.[11] For gender bias in coreference and hate speech, we use terms that are male vs. female and are career, math, science, vs. family, art. For gender bias and migrant bias in Spanish hate speech, we compare male/female identity or migrant/non-migrant identity with pleasant-unpleasant term expansions.[12]

[7] This fact is the premise behind the work of Stanovsky et al. (2019) who use the explicit marking in translation to reveal bias.

[8] Details of datasets & preprocessing are in Appendix C.

[9] WEAT uses wordlists to measure relationships between words in the space, and bias modification depends on identifying words to sub or supersample (for databalancing), or to adjust (for Attract-Repel). Many other debiasing methods that we did not use (e.g Bolukbasi et al. (2016)) also use wordlists.

[10] All WEAT wordlists are in Appendix B. We make a small substitution of general gender words instead of proper names in WEAT 6, as proper names by design do not appear in our coreference task.

[11] Final word sets are 200-400 words, due to significant overlap in nearest neighbors & manual removal of odd terms.

[12] We did additionally experiment with using the *exact*

4.3 New Spanish WEAT

We substantially modified Spanish WEAT (aka XWEAT for non-English WEATs) and added entirely new terms. The reason for this is that the original XWEAT was translated from English very literally, which causes two problems.

The first problem with XWEAT is that many of the terms do not make sense in a Spanish speaking community — names included in the original, like *Amy*, are names in Spanish and thus were untranslated, but are uncommon and have upper class connotations not intended in the original test. Another example is *firearms* translated as *arma de fuego*, which while technically a correct literal translation, is not commonly used to describe weapons.[13]

The second problem with XWEAT is that nouns on the wordlists for both abstract math and science concepts as well as abstract art concepts are almost entirely grammatically female. For instance, *ciencia* (science), *geometría* (geometry) are grammatically female, as are *escultura* (sculpture) and *novela* (novel). It is well established that for languages with grammatical gender, words that share a grammatical gender have embeddings that are closer together than words that do not (Gonen et al., 2019; McCurdy and Serbetci, 2017). So, when WEAT in English was translated into XWEAT in Spanish (Glavas et al., 2019), the terms were imbalanced with regard to grammatical gender, which makes the results misleading. We balance the lists, often replacing abstract nouns with corresponding adjectives which can take male or female form, e.g. *científico* and *científica* (scientific, male and female), such that we can use both versions to account for the effect of grammatical gender.

Finally, we needed a metric to examine bias against migrants. Metrics for intrinsic bias must be targeted to the type of harm expected in the downstream application, and there is not an out-of-the-box WEAT test for this. So we create a new WEAT test for bias against migrants in Spanish. Following the setup of tests for racial bias in original WEAT — based on American racial biases in English — we have lists of names associated with migrants vs. non-migrants, and compare them with lists of pleasant and unpleasant terms. The names are based on work of Salamanca and Pereira (2013), who studied ranking names as lower vs. upper class; class status is closely correlated with whether a person is a migrant. We select a subset of names in which the majority in the study agree on the class. Pleasant and unpleasant terms exist in WEAT and XWEAT, but we again modify them to balance grammatical gender.

5 Results

Figure 3 displays data for all tasks: one scatterplot per triple of experimental variables: an intrinsic metric, an extrinsic metric, an embedding algorithm. If we want to be able to broadly use WEAT metrics for any given bias research, these graphs should each show a clear and a positive correlation. None of them do. There are no trends in correlation between the metrics that hold in all cases regardless of experimental detail, for any of the tasks. We have additionally examined whether there are correlations within one bias modification method (pre or postprocessing) in case a difference in the way these methods modify embeddings causes differences in trends. In most cases this breakout tells the same story. The select cases where positive (and negative) correlations are present are discussed below. Further breakout graphs and combinations are included in Appendix D.

Coreference (en): Gender The coreference task (Figure 3, rows 1-3) doesn't display a clear correlation in all cases, and yet it has the clearest relationship of all three tasks, with a significant moderate positive correlation for both Predictive Parity (precision) and Equality of Opportunity (recall) for word2vec (columns 3 & 4). The overall trends are muddied by the data for fastText, which does not have a significant correlation under any conditions. Both are expected: that coreference would display the strongest trends, and that fastText would display more unpredictable or weaker trends. The Winobias coreference task is as directly matched to the WEAT tests as it is possible to be - since both use common career words to measure bias. So the relationship between the two metrics is clearest here: moving female terms closer to certain career terms most directly helps a system resolve anti-stereotypical coreference chains. However, we still only see a correlation in wod2vec, not fastText. fastText may behave less predictably because of its use of subwords; when subwords are used,

WEAT terms for debiasing, and found the trends to be similar but of smaller magnitude, so we settled on expanded lists as a more realistic scenario.

[13] The standard would be *armas*. *arma de fuego* is also composed of three words, and so will not appear in any vocabulary.

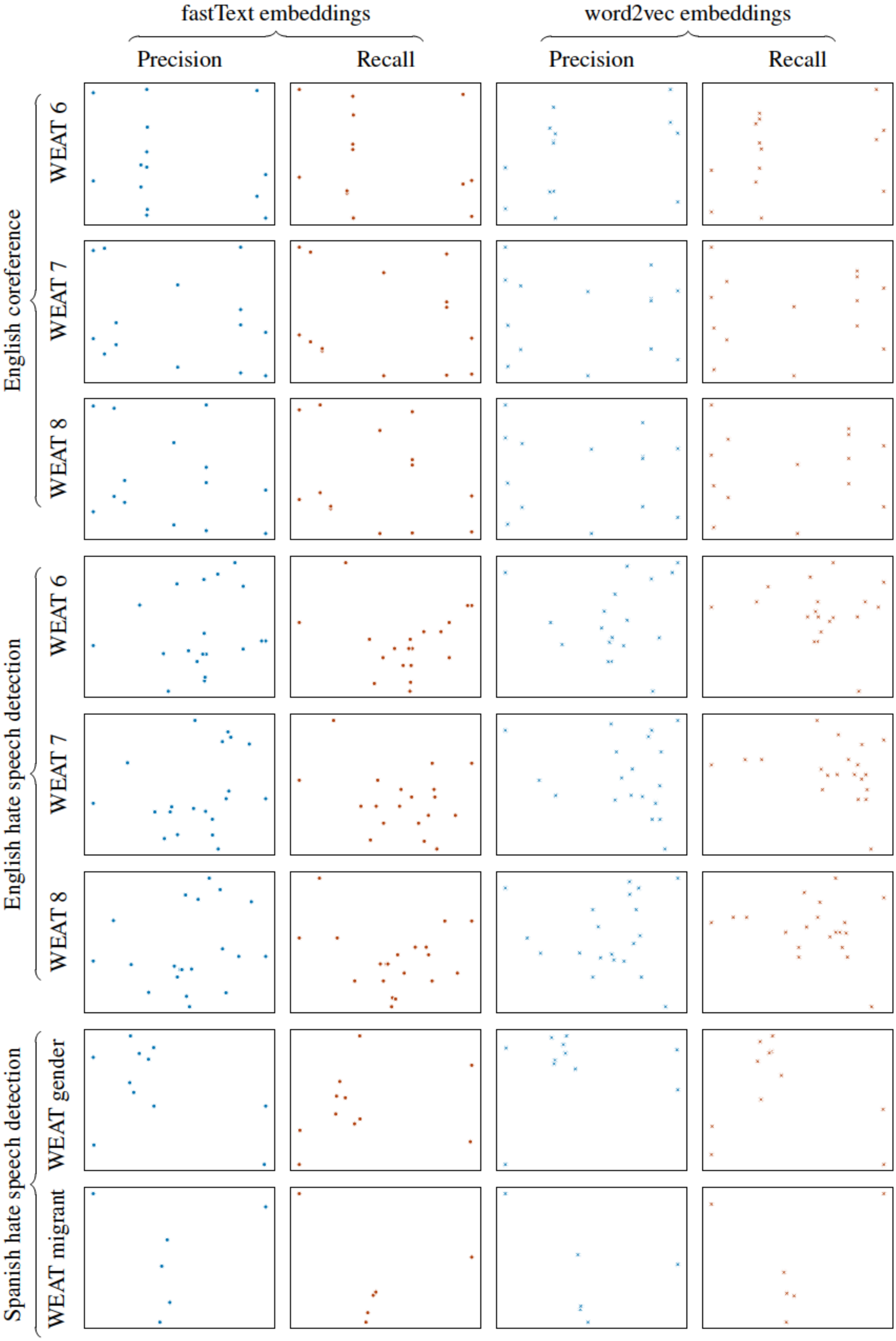


Figure 3: Experimental results, showing one scatterplot per experiment. An experiment consists of a task (outer row label), an embedding (outer column label), an intrinsic metric (inner row label), and an extrinsic metric (inner column label). Each point in a scatterplot is the intrinsic (y-axis) and extrinsic (x-axis) measure of bias for a single run, where word embeddings for each run have been debiased or overbiased using pre- or post-processing.

word representations are more interconnected.[14] We can debias with regard to a specific word, but that word's embedding will still be influenced by all other words that share its character ngrams. It is difficult to predict how changing the composition of a training corpus will affect all words that contain a certain ngram (e.g. *ch*) in them. For this reason, fastText may be initially more resistant to encoding biases than word2vec, as was found in Lauscher and Glavas (2019), but may also be more complex to debias. This has implications for extending this work to contextual models, which always use some form of subword unit.

Hatespeech (en): Gender Hatespeech (en) has fewer and more restricted correlations than coreference, as can be seen in Figure 3, rows 4-6. These plots show no relationship at all between intrinsic and extrinsic metrics. When data is broken out by bias modification method (see Figure 4b in Appendix D), it becomes clear that there is a moderate *positive* correlation for postprocessing for recall, and the aggregate appears this way because there is a moderate *negative* correlation for preprocessing. This holds for both embedding algorithms, though both positive and negative correlations are stronger for fastText. Precision displays no correlation. Note that the absolute variance in recall is much smaller than for precision, but this is still significant for each embedding algorithm individually and for both grouped together.

Of interest for future bias research is that the baseline level of bias (premodification, from raw twitter data) in English hatespeech differs by embedding type, but *only* for precision. Initial models (with unmodified embeddings) using fastText have 10 additional points of precision for male-targeted hatespeech than for female-targeted. However initial models using word2vec have the opposite bias and have 4 fewer points of precision for male-targeted than female targeted hatespeech. For recall, the two embedding algorithms are equivalent, with 6 fewer points for male-targeted hatespeech. In fact, in the recall metric there is an early indication of unreliability of the relationship we are examining between WEAT and extrinsic bias, because there is a spread of different WEAT results that map to nearly the same difference in recall.

Hatespeech (es): Gender and Migrant For hatespeech in Spanish, we examine two kinds of bias separately — gender bias and bias against migrants, in Figure 3, rows 7,8. Neither gender bias nor migrant bias show positive correlations in any experimental conditions.

Gender bias in our models is in an *absolute* sense never present, since in Spanish hatespeech targeted against women is easier to identify than against others (with F1 in the high 80s).[15] But there are no overall trends when this is bias is modified to be more or less extreme, and there are no positive correlations in any conditions. There is a moderate *negative* correlation for precision only when looking at fastText embeddings.

Migrant bias similarly has no trends save in very restricted conditions broken out by bias modification type. In contrast to the gender case, hatespeech against migrants is clearly challenging to identify, with much lower F1 in the low 60s. There is a positive correlation between migrant bias and performance gap for recall with preprocessing in fastText only. This fits the expectation that fastText may be more sensitive to preprocessing than postprocessing due to subwords, as discussed above, though in the gender bias case with negative correlation it is equally sensitive to both, so it is hard to draw conclusions. Given the smaller number of datapoints for Spanish (discussed below) this is likely just noise. To confuse the situation further, the only trends in precision are present in word2vec, and are negative correlations.

Note that all graphs for Spanish display central clusters, because it was more difficult to get an even spread of bias measures, and because Spanish has fewer data points than English. This is for a number of reasons that compound and underscore the difficulty of expanding supposedly language-agnostic techniques beyond English, even to high resourced languages like Spanish. We have only one WEAT test for each type of bias, since we made our own that carefully balanced grammatical gender, after rectifying the issues with the existing translated versions (see Section 4.3). Bias modification is also more difficult - the richer agreement system in Spanish means that there are more surface forms of what would be one word in English. In addition, the language model used for nearest neighbour expansion of wordlists (see Sec-

[14] For example, the representation of the word *childish* is by design also made up of the representations for *child* and *ish*, but also all unigrams, bigrams, and trigrams it contains (*c*, *ch*, *chi*, etc).

[15] This is perhaps due to examples in the training data having clearer markers such as specific anti-female slurs, but is itself an interesting question.

tion 4.2) produces predominantly formal register words from news or scientific articles, due to a less varied makeup of its training data than the English model. This makes them less well suited to debiasing twitter data specifically, and there were no readily available models that had more casual register. For bias against migrants, there is the additional challenge that wordlists are predominantly based on proper names, which are much rarer in twitter (which tends to use @ mentions instead) than in other media.

6 Discussion

The broad result of this research is that changes in WEAT do not correlate with changes in application bias, and therefore that WEAT should not be used to measure progress in debiasing algorithms. We have found that even when we maximally target bias modification of an embedding, we cannot produce a reliable change in bias direction downstream. There was no pattern or correlation between tasks, for the same task in different languages, or even in most cases within one task. And we have chosen one of the simplest possible setups, with full-word embeddings and a single type of bias at a time. Real world scenarios can easily be more complicated and involve multiple types of bias or subword embeddings. Our findings also indicate that additional complexity may muddy the relationship further. For example, fastText behaved less predictably than word2vec across experiments, suggesting that if we were to expand to larger models that are fully reliant on subwords the patterns may become even less clear.

The implication of this finding is that an NLP scientist or engineer has limited options when investigating and mitigating bias. They must a) find the specific set of wordlists, embedding algorithms, downstream tasks, and bias modification methods that are together predictive of bias for the given task, language, and model or b) implement full systems to test application bias directly, even if their work focuses on embeddings.

While the latter may seem onerous, it may not be more so than exhaustively searching for a configuration where intrinsic bias metrics are predictive.

This underscores the importance of making good downstream bias measures available, as either approach will require these. More datasets that are collected need to be annotated with subgroup demographic and identity information — there are very few available. More research needs to focus on creating good challenge sets to measure application bias. Additional research on more broad usage of unsupervised methods (Zhao and Chang, 2020) would also be valuable, though those also would benefit from subgroup identity annotation to make their results more interpretable.

It is only when more of these things are readily available that we can see the true measure of the efficacy of our debiasing efforts.

We do note a limitation of this study in that all downstream tasks are discriminative classification tasks. Bias in classification is more straightforward to measure, with well established metrics, but covers allocational harms (performance disparity), whereas the inclusion of generative models could better cover representational harms (misleading or harmful representations/portrayals) (Blodgett et al., 2020; Crawford, 2017). Concurrent research on causal mediation analysis for bias has shown that the embedding layer in open-domain generation has the strongest effect on gender bias (as compared to other layers of the network) (Vig et al., 2020). Further work could investigate whether generation tasks have display the same or different relationship to intrinsic metrics.

7 Conclusion

We have examined the relationship of the intrinsic bias metric WEAT to the extrinsic bias metrics of Equality of Opportunity and Predictive Parity, for multiple tasks and languages, and determined that positive correlations between them exist only in very restricted settings. In many cases there is either negative correlation or none at all. While intrinsic metrics such as WEAT remain good descriptive metrics for computational social science, and for examining bias in human texts, we advise that the NLP community not rely on them for measuring model bias. We instead advise that they focus on careful consideration of downstream applications and the creation of datasets and challenge sets that enable measurement at this stage.

Acknowledgements

We thank Andreas Grivas, Kate McCurdy, Yevgen Matusevych, Elizabeth Nielsen, Ramon Sanabria, Ida Szubert, Sabine Weber, Björn Ross, Agostina Calabrese, and Eddie Ungless for comments on earlier drafts of this paper.

References

Valerio Basile, C. Bosco, E. Fersini, Debora Nozza, V. Patti, F. Pardo, P. Rosso, and M. Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *SemEval@NAACL-HLT*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186.

Kate Crawford. 2017. The trouble with bias. (keynote at neurips).

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Scott Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *AIES '18*.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *ICWSM*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *EMNLP*.

Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate crosslingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *ACL*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NAACL-HLT*.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 463–471, Hong Kong, China. Association for Computational Linguistics.

A. Greenwald, D. McGhee, and J. L. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74 6:1464–80.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *NIPS*.

Felix Hill, Roi Reichart, and A. Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41:665–695.

Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un)fairness: Lessons for machine learning. In *FAT* '19*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Keita Kurita, N. Vyas, Ayush Pareek, A. Black, and Yulia Tsvetkov. 2019. Quantifying social biases in contextual word representations. In *1st ACL Workshop on Gender Bias for Natural Language Processing*.

Anne Lauscher and Goran Glavas. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In **SEM@NAACL-HLT*.

Anne Lauscher, Goran Glavas, Simone Paolo Ponzetto, and Ivan Vulic. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *AAAI*.

Kenton Lee, Luheng He, M. Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *EMNLP*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *NAACL-HLT*.

K. McCurdy and Oguz Serbetci. 2017. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *WiNLP*.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Nikola Mrksic, Ivan Vulic, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gasic, Anna Korhonen, and Steve J. Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Gastã Salamanca and Lidia Pereira. 2013. PRESTIGIO Y ESTIGMATIZACIÃ“N DE 60 NOMBRES PROPIOS EN 40 SUJETOS DE NIVEL EDUCACIONAL SUPERIOR. *Universum (Talca)*, 28:35 – 57.

Maarten Sap, D. Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.

João Sedoc and Lyle Ungar. 2019. The role of protected class word lists in bias identification of contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 55–61, Florence, Italy. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *ACL*.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *EthNLP@EACL*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

R. Weischedel, E. Hovy, M. Marcus, and Martha Palmer. 2017. Ontonotes : A large training corpus for enhanced processing.

Jieyu Zhao and Kai-Wei Chang. 2020. LOGAN: Local group bias detection by clustering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1968–1977, Online. Association for Computational Linguistics.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *ACL*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL-HLT*.

Jieyu Zhao, Yichao Zhou, Z. Li, W. Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *EMNLP*.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *EMNLP/IJCNLP*.

A Bias Metric Definitions & Formulas

Performance Gap metrics measure difference in performance across different demographic splits of the data, and are in our case (and most commonly) applied to classification tasks.

Where A is a demographic variable (race, gender, etc), Y is the true label, and $\hat{Y}$ is the predicted label, a fair system will satisfy:

$$P(\hat{Y} = 1|A = x, Y = 1) = P(\hat{Y} = 1|A = y, Y = 1)$$

where x and y are demographic values usually of an *privileged* and a *underprivileged* group. This expresses that the probability of a given test sample being correctly identified as a true positive should be equal regardless of group, and is known as **Equality of Opportunity** (Hardt et al., 2016). A fair system will also satisfy:

$$P(\hat{Y} = 1|A = x, Y = 0) = P(\hat{Y} = 1|A = y, Y = 0)$$

which expresses that that probability of a given test sample being incorrectly identified as positive is equal regardless of group. This is known as **Predictive Parity** and when combined with Equality of Opportunity is known as **Equalized Odds**.

These are easily measured in most NLP systems. The former is captured by measuring recall gap, where if x is the privileged group and y the underprivileged, unfairness is captured by $Recall_x - Recall_y$, where any positive value is unfair. The latter is captured by $Precision_x - Precision_y$, again where positive values are unfair.

B WEAT Formula and Wordlists

B.1 English WEAT lists

All are tests for gender bias.

B.1.1 Weat 6

WEAT 6 was modified to use the general gender terms of 7,8 rather than proper names, because the co-reference task contains no names.

Male *male, man, boy, brother, he, him, his, son*

Female *female, woman, girl, sister, she, her, hers, daughter*

Career *executive, management, professional, corporation, salary, office, business, career*

Family *home, parents, children, family, cousins, marriage, wedding, relatives*

The original WEAT 6 uses the following male and female names as the gender terms:
Male: *John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill*
Female: *Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna*.

B.1.2 Weat 7

Male *male, man, boy, brother, he, him, his, son*

Female *female, woman, girl, sister, she, her, hers, daughter*

Math *math, algebra, geometry, calculus, equations, computation, numbers, addition*

Art *poetry, art, dance, literature, novel, symphony, drama, sculpture*

B.1.3 Weat 8

Male *brother, father, uncle, grandfather, son, he, his, him*

Female *sister, mother, aunt, grandmother, daughter, she, hers, her*

Science *science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy*

Art *poetry, art, Shakespeare, dance, literature, novel, symphony, drama*

B.2 Changes to English List

We modify WEAT 6 to use the gender terms for WEAT 7/8 as the terms for 6, but otherwise leave terms as is.

WEAT 6 (career/family vs. male/female) uses proper names as gender terms, whereas the other two tests use more standard gender terms (she, her, he, him, mother, father). This is an artifact of replicating IAT, which introduces a confound in their comparability – if the WEAT tests have different patterns of correlation, we don't know whether this is because of the difference in the way gender bias patterns for career/family vs. for arts/science or whether it patterns differently because of proper names vs. gender terms. This is exacerbated in our case where proper names are treated even more differently than usual both in twitter (where @mentions stand in for proper names) and in the Winobias metric that we use (where professions are used instead of proper names precisely because names contain gender information and the challenge set intends to be ambiguous).

B.3 Spanish WEAT lists:

Recall that we created these ourselves, the gender test with reference to both the original gender focused WEAT 6,7,8 of Caliskan et al. (2017) and the translation of Lauscher and Glavas (2019), significantly modified and extended to balance grammatical gender across sets of words. The migrant test was created with reference to the tests for racism that use African-American vs. European-American names paired with pleasant vs. unpleasant terms in WEAT 3, 4, 5, using the lists of European Spanish vs. migrant Spanish names identified by Salamanca and Pereira (2013).

B.3.1 Gender

Male: *masculino, hombre, niño, hermano, él, hijo, hermano, padre, papá, tío, abuelo*

Female: *femenino, mujer, niña, hermana, ella, hija, hermana, madre, mamá, tía, abuela*

Science: *científico, físico, químico, astrónomo, tecnológico, biólogo, científica, física, química, astrónoma, tecnológica, bióloga*

Art: *arquitecto, escultor, pintor, escritor, poeta, bailarín, actor, fotógrafo, arquitecta, escultora, pintora, escritora, poetisa, bailarina, actora, fotógrafa*

B.3.2 Migrants

European-Spanish names: *Agustina, Martina, Josefa, Antonia, Sofía, Isidora, Cristóbal, Sebastián, Agustín, Alonso, Joaquín, León, Ignacio, Julieta, Matilde*

Migrant-Spanish names: *Shirley, Yamileth, Sharon, Britney, Maryori, Melody, Nayareth, Yaritza, Byron, Brian, Jason, Malcon, Justin, Jeremy, Jordan, Brayan, Yeison, Yeremi, Bairon, Yastin*

Pleasant terms: *caricia, libertad, salud, amor, paz, animar, amistad, cielo, lealtad, placer, diamante, gentil, honestidad, suerte, arcoiris, diploma, regalo, honor, milagro, amanecer, familia, alegría, felicidad, risa, paraíso, vacación, paz, maravilloso, maravillosa*

Unpleasant terms: *abuso, choque, suciedad, asesinato, enfermedad, accidente, muerte, sufrimiento, veneno, hedor, apestar, ataque, asalto, desastre, odio, contaminación, tragedia, divorcio, cárcel, pobreza, fea, feo, cáncer, matar, vómito, bomba, maldad, podrido, podrida, agonía, terrible, horrible, guerra, repugnante*

C Training Data and Preprocessing

This details the data for training embeddings. For data used in training the final models, see relevant papers cited in Section 4.1.

C.1 Wikipedia

Wikipedia data is downloaded from the latest Wikipedia article dump, tokenized with NLTK (`https://www.nltk.org/`), and all words appearing less than 10 times are replaced with `<unk>`. The final dataset has 439,935,872 words.

C.2 Twitter

Twitter data is from 2019 and is downloaded from the Internet Archive `https://archive.org/details/twitterstream`. Retweets are removed, and data is lowercased, tokenized with NLTK TweetTokenizer, and hashtags and @mentions are replaced with `<HASH>` and `<MENTION>` respectively. All words appearing less than 10 times are replaced with `<unk>`. English twitter data size is 3,641,306 tweets with 38,376,060 words. Spanish twitter data size is 10,683,846 tweets with 142,715,339 words.

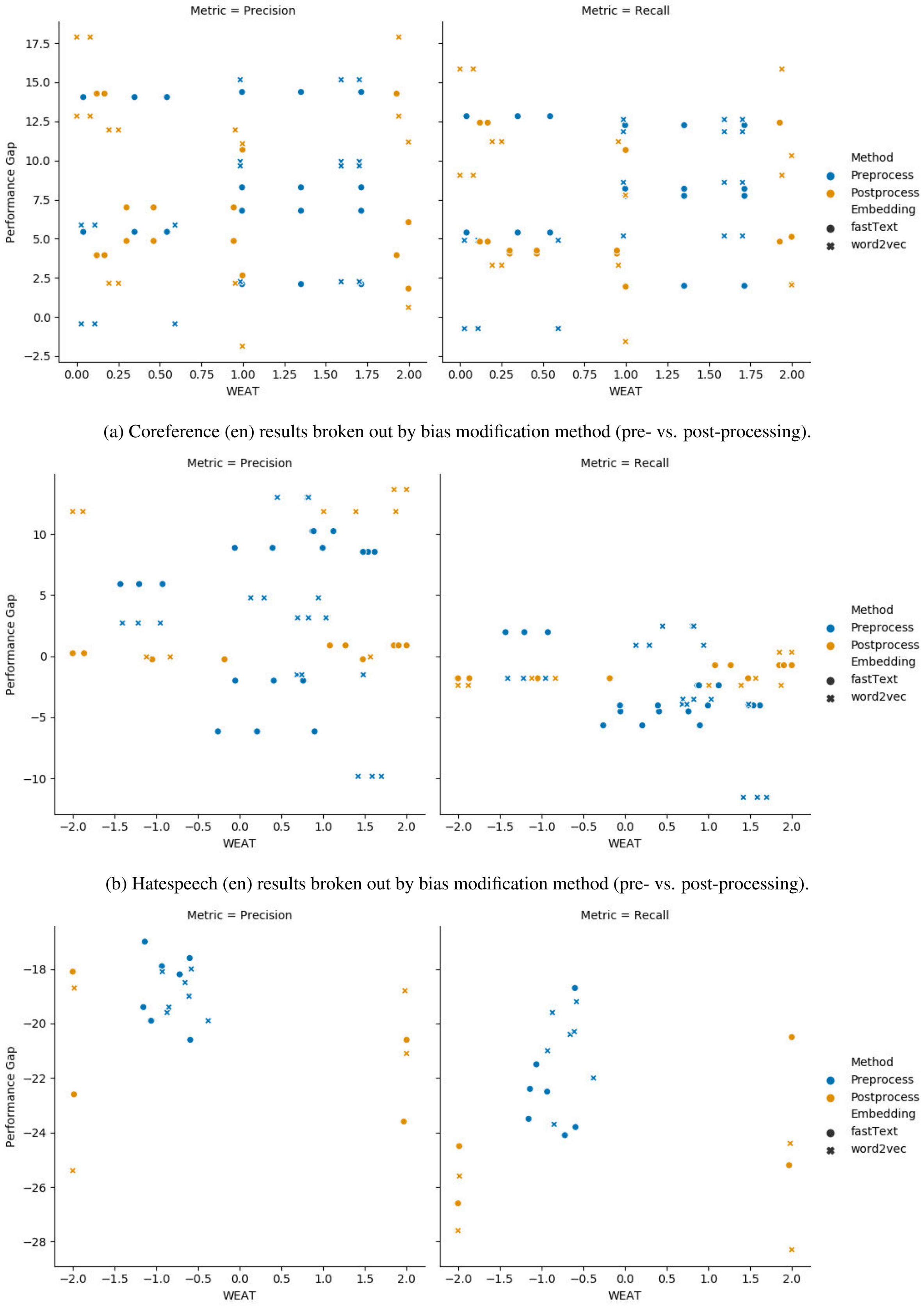
D Further Results Graphs

Below are breakouts of graphs by bias modification method, as well as full graphs with metric scales and legends.

Figure 4 breaks out all tasks by bias modification method (pre- vs. post-processing). The main interesting thing to note here is for hatespeech in English. Based on the spread of data points, it is easy to see that there is overall more effect on precision gap when embeddings are modified, whereas recall performance gap occupies a narrower band over a wide spread of WEAT metrics. Yet recall is the only metric which has a positive correlation with WEAT, and then only in the postprocessing condition. For Spanish it is also visible that it is much more difficult to modify bias for Spanish when preprocessing vs. when postprocessing.

Figure 5 shows one graph for each task and bias type combination, in full, in order to view the effect of not controlling for experimental variable. It also shows the scale for the spread of data points.

Finally, for interest, we also include Figure 6, which displays the correlation broken out by type of Winobias test (which differ in difficulty because Type 1 is semantic and Type 2 is syntactic).

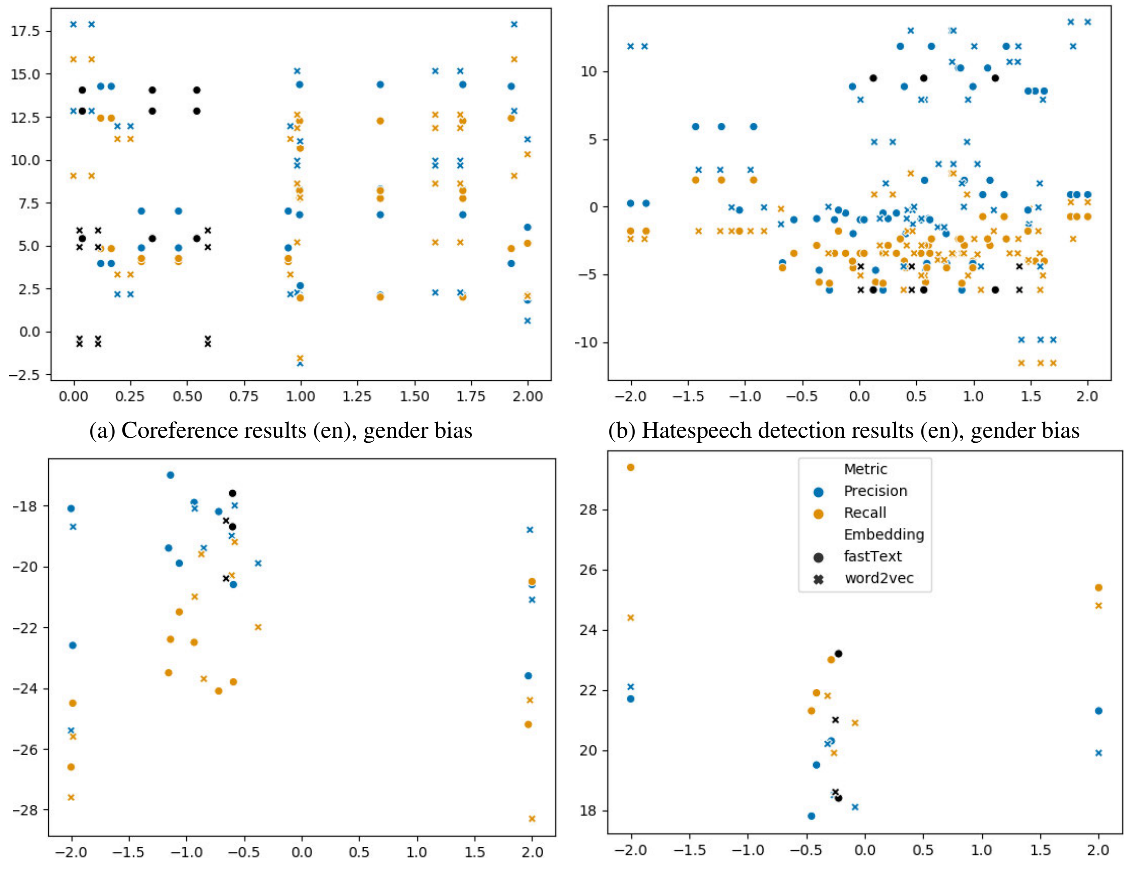


(a) Coreference (en) results broken out by bias modification method (pre- vs. post-processing).

(b) Hatespeech (en) results broken out by bias modification method (pre- vs. post-processing).

(c) Hatespeech (es) results for gender bias metrics broken out by bias modification method.

Figure 4: Bias modification method breakout by pre vs. post-processing for gender bias for each task for both precision and recall.



(a) Coreference results (en), gender bias

(b) Hatespeech detection results (en), gender bias

(c) Hatespeech detection results (es), gender bias

(d) Hatespeech detection results (es), migrant bias

Figure 5: Scatterplots showing all data points for each of the 4 tasks: gender bias in co-reference (en), gender bias in hatespeech detection (en), gender bias in hatespeech detection (es), and migrant bias in hatespeech detection (es). In each plot, the x-axis represents WEAT, and the y-axis shows performance gap between groups (male-female, female-other, migrant-other). Original embeddings (before modification) shown in black. There is no correlation that holds independently of experimental conditions (embedding type, bias modification method, WEAT test).

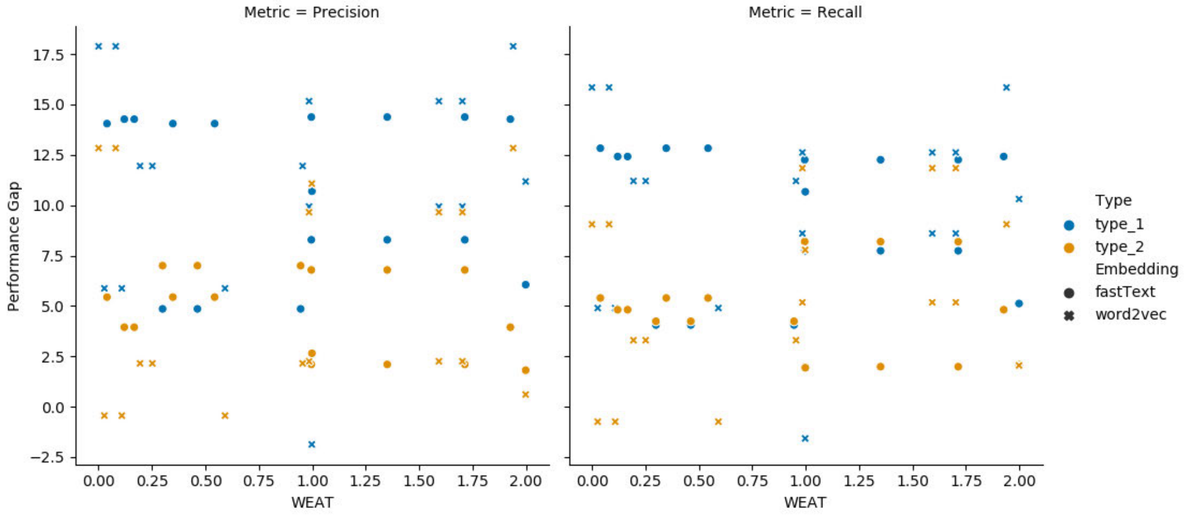


Figure 6: Coreference (en) results broken out by type of Winobias challenge, Type 1 is more difficult as there are only semantic cues to correct coreference, Type 2 has also syntactic cues.

Chapter 4

How Gender Debiasing Affects Internal Model Representations, and Why It Matters

The previous work showed that WEAT, the common measure of bias in embedding spaces, doesn't correlate with application bias. Debiasing at the language model could still sometimes work, but we showed that we cannot tell if it has worked without implementing a downstream system. We recommended that bias be always tested in a downstream application.

This recommendation was strengthened when our work was later replicated with contextual embeddings in Cao et al. (2022), who study a less extensive set of demographics and of languages, but a more broad set of intrinsic and extrinsic metrics, for 19 different contextualised models. They still find no reliable correlation, thought they make small modifications to intrinsic and extrinsic metrics to try to make them align better.

But implementing downstream systems is exactly what the field is trying to avoid. The increase in scale of pre-training over the past five years is only exacerbating this; less and less pre-training is done by people who deploy systems. I pre-trained the embeddings in the previous chapter (3) on a university cluster, but very few train new BERT models, and only a handful train LLMs.

So in the following, we make progress on understanding the relationship between intrinsic and extrinsic bias by studying the reverse direction. We cannot yet tell in what way modifying a language model representation affects downstream bias, so how does downstream debiasing affect an upstreadm langauge model? We know a priori that debiasing at the second stage of transfer learning works. So perhaps this is a better place to start, and it will be more enlightening to look at how debiasing downstream (the thing we understand better) affects representations upstream (the thing we understand less well).

This motivates the following work, we try to understand language model representations better by studying the impact of downstream debiasing. We know that downstream debiasing does improve downstream bias metrics. If the language model is not 'frozen' (e.g. the downstream debiasing backpropagates to change the language model parameters) then this changes model representations as well. This worked. We found the CEAT metric (contextualised WEAT) to be as uncorrelated in this reverse direction of downstream task → language model as we and (Cao et al., 2022) had found WEAT and CEAT to be when going in the original direction of language model → downstream task. But we found that **information theoretic probing** could be adopted as a good gender bias metric when applied to gender demographic information. Information theoretic probing had previously been used as a method of analysis of compressibility of linguistic properties of a learned representation: POS tags, dependency parses, etc. By this 'reversed' method of the previous analysis we were able to adapt it into an upstream language model bias metric.

In this work we also disentangled of the role of the language model vs. the downstream classifier for fairness in transfer learning. Through information theoretic probing we were able to identify a language model's 'potential' for gender bias, which then may or may not be realised by the classifier depending on the downstream fine-tuning data. When Sandra Kublik interviewed me and we discussed this work, she suggested a genetic analogy to give intuition for this behaviour clearly to laypeople. I can have a genetic propensity for breast cancer, or for schizophrenia, but that may or may not ever be realised depending on my environmental factors. Similarly, if the language model has strong *potential* for gender bias, and the downstream fine-tuning data is imbalanced such that gender A is a strong predictor of labels Y, then the language model will be biased. But even with strong potential, if the fine-tuning data is not imbalanced, then the potential will not be realised. Conversely, with a language model with lower potential, the imbalance in fine-tuning data has a smaller effect. So this work shows that, just as with genetics, the full story cannot be determined from the language model, but some of the story can be.

In this work we also used the full suite of ten bias metrics that are generally applied to classification, which is rarely done. They tend to track with each other, but for one of the two tasks had very different magnitudes, so had we studied a smaller subset of them, we may have come to different conclusions. Fairness metrics must be matched to downstream applications, as discussed in §2.2, but for work that analyses general relationships between metrics, we show that it is necessary to include the broad range to draw robust conclusions. We also find a pattern in a specific type of bias metric: Pearson correlation to real-word bias statistics, and show it to be flawed. Pearson correlation metrics correlate one of the performance gap metrics mentioned in §2.2 to real world disparities, which for these studies is the gender bias in profession classification or profession based co-reference resolution to gender disparities in professions in the real-world (usually the United States). We show that these metrics are extremely unreliable because they hide a confound: whether or not the statistics of the pre-training data reflect the statistics of the real world. So we recommend at least using one of the other metrics, unless the relationship to the real world is the object of study itself, rather than the bias behaviour of a language model (this is rarely the case in work in ML venues). I did not include correlation based metrics in my overview of fairness metrics §2.2 because of these flaws.

How Gender Debiasing Affects Internal Model Representations, and Why It Matters

Hadas Orgad[1] **Seraphina Goldfarb-Tarrant**[2] **Yonatan Belinkov**[1]*

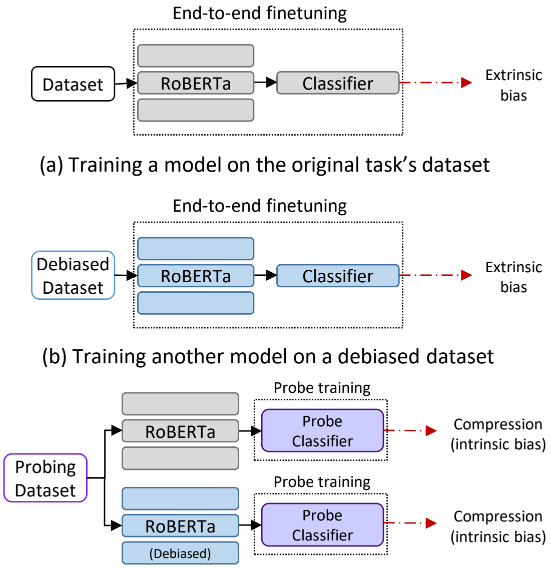
[1]Technion – Israel Institute of Technology [2]University of Edinburgh

Abstract

Common studies of gender bias in NLP focus either on extrinsic bias measured by model performance on a downstream task or on intrinsic bias found in models' internal representations. However, the relationship between extrinsic and intrinsic bias is relatively unknown. In this work, we illuminate this relationship by measuring both quantities together: we debias a model during downstream fine-tuning, which reduces extrinsic bias, and measure the effect on intrinsic bias, which is operationalized as bias extractability with information-theoretic probing. Through experiments on two tasks and multiple bias metrics, we show that our intrinsic bias metric is a better indicator of debiasing than (a contextual adaptation of) the standard WEAT metric, and can also expose cases of superficial debiasing. Our framework provides a comprehensive perspective on bias in NLP models, which can be applied to deploy NLP systems in a more informed manner. [1]

1 Introduction

Efforts to identify and mitigate gender bias in Natural Language Processing (NLP) systems typically target one of two notions of bias. *Extrinsic* evaluation methods and debiasing techniques focus on the bias reflected in a downstream task (De-Arteaga et al., 2019; Zhao et al., 2018), while *intrinsic* methods focus on a model's internal representations, such as word or sentence embedding geometry (Caliskan et al., 2017; Bolukbasi et al., 2016; Guo and Caliskan, 2021). Despite an abundance of evidence pointing towards gender bias in pretrained language models (LMs), the extent of harm caused by these biases is not clear when it is not reflected in a specific downstream task (Barocas et al., 2017; Kate Crawford, 2017; Blodgett et al., 2020; Bommasani et al., 2021). For instance, while the word embedding proximity of "doctor" to "man" and "nurse" to "woman" is intuitively normatively wrong, it is not clear when such phenomena would lead to downstream predictions manifesting in social biases. Recently, Goldfarb-Tarrant et al. (2021) have shown that debiasing static embeddings intrinsically is not correlated with extrinsic gender bias measures, but the nature of the reverse relationship is unknown: how are extrinsic interventions reflected in intrinsic representations? Furthermore, Gonen and Goldberg (2019a) demonstrated that a number of intrinsic debiasing methods applied to static embeddings only partially remove the bias and that most of it is still hidden within the embed-



(c) Measuring our intrinsic metric on the debiased and original dataset using probing

Figure 1: Our proposed framework. Black arrows mark forward passes, red arrows mark things we measure. We first (a) train a model on a downstream task, then (b) train another model on the same task using a debiased dataset, and finally (c) measure intrinsic bias in both models and compare.

[*] Supported by the Viterbi Fellowship in the Center for Computer Engineering at the Technion.

[1] Our code and model checkpoints are publicly available at https://github.com/technion-cs-nlp/gender_internal

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2602 - 2628
July 10-15, 2022 ©2022 Association for Computational Linguistics

ding. Complementing their view, we examine *extrinsic* debiasing methods, as well as demonstrate the possible harm this could cause. Contrary to their conclusion, we do not claim that these debiasing methods should not be trusted, *as long as they are utilized with care*.

Our goal is to gain a better understanding of the relationship between a model's internal representations and its extrinsic gender bias by examining the effects of various debiasing methods on the model's representations. Specifically, we fine-tune models with and without gender debiasing strategies, evaluate their external bias using various bias metrics, and measure intrinsic bias in the representations. We operationalize intrinsic bias via two metrics: First, we use CEAT (Guo and Caliskan, 2021), a contextual adaptation of the widely used intrinsic bias metric WEAT (Caliskan et al., 2017). Second, we propose to use an information-theoretic probe to quantify the degree to which gender can be extracted from the internal model representations. Then, we examine how these intrinsic metrics correlate with a variety of extrinsic bias metrics that we measure on the model's downstream performance. Our approach is visualised in Figure 1.

We perform extensive experiments on two downstream tasks (occupation prediction and coreference resolution); several debiasing strategies that involve alterations to the training dataset (such as removing names and gender indicators, or balancing the data by oversampling or downsampling); and a multitude of extrinsic bias metrics. Our analysis reveals new insights into the way language models encode and use information on gender:

- The effect of debiasing on internal representations is reflected in gender extractability, while not always in CEAT. Thus, gender extractability is a more reliable indicator of gender bias in NLP models.
- In cases of high gender extractability but low extrinsic bias metrics, the debiasing is superficial, and the internal representations are a good indicator for this: The bias is still present in internal representations and can be restored by retraining the classification layer. Therefore, our proposed measuring method can help in detecting such cases before deploying the model.
- The two tasks show different patterns of correlation between intrinsic and extrinsic bias. The coreference task exhibits a high correlation. The occupation prediction task exhibits a lower correlation, but it increases after retraining (a case of superficial debiasing). Gender extractability shows higher correlations with extrinsic metrics than CEAT, increasing the confidence in this metric as a reliable measure for gender bias in NLP models.

2 Methodology

In this study, we investigate the relationship between extrinsic bias metrics of a task and a model's internal representations, under various debiasing conditions, for two datasets in English. We perform extrinsic debiasing, evaluate various extrinsic and intrinsic bias metrics before and after debiasing, and examine correlations.

Dataset. Let $D = \{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}\}$ be a dataset consisting of input data $\boldsymbol{X}$, labels $\boldsymbol{Y}$ and protected attributes $\boldsymbol{Z}$.[2] This work focuses on gender as the protected attribute z. In all definitions, F and M indicate female and male gender, respectively, as the value of the protected attribute z.

Trained Model. The model is optimized to solve the downstream task posed by the dataset. It can be formalized as $f \circ g : \boldsymbol{X} \to \mathbb{R}^{|\mathcal{Y}|}$, where $g(\cdot)$ is the feature extractor, implemented by a language model, e.g., RoBERTa (Liu et al., 2019), $f(\cdot)$ is the classification function, and $\mathcal{Y}$ is the set of the possible labels for the task.

2.1 Bias Metrics

Each bias evaluation method described in the literature can be categorized as extrinsic or intrinsic. In all definitions, $\mathbf{r}$ indicates the model's output probabilities.

2.1.1 Extrinsic Metrics

Extrinsic methods involve measuring the bias of a model solving a downstream problem. The extrinsic metric is a function:

$$E(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{R}, \boldsymbol{Z}) \in \mathbb{R}$$

The output represents the quantity of bias measured; the further from 0 the number is, the larger the bias is. Our analysis comprises a wide range

[2] $\boldsymbol{Z}$ is by convention used for attributes for which we want to ensure fairness, such as gender, race, etc. It is purposefully broad, and depending on the task and data could refer to the gender of an entity in coreference, the subject of a text, the demographics of the author of a text, etc.

of extrinsic metrics, including some that have been measured in the past on the analyzed tasks (Zhao et al., 2018; De-Arteaga et al., 2019; Ravfogel et al., 2020; Goldfarb-Tarrant et al., 2021) and some that have never been measured before, and shows our results apply to many of them. For illustration, we will consider occupation prediction, a common task in research on gender bias (De-Arteaga et al., 2019; Ravfogel et al., 2020; Romanov et al., 2019). The input x is a biography and the prediction y is the profession of the person described in it. The protected attribute z is the gender of that person.

Performance gap. This is the difference in performance metric for two different groups, for instance two groups of binary genders, or a group of pro-stereotypical and a group of anti-stereotypical examples. We measure the following metrics: True Positive Rate (TPR), False Positive Rate (FPR), and Precision. In occupation prediction, for instance, the TPR gap for each profession y expresses the difference in the percentage of women and men whose profession is y and are correctly classified as such. We also measure F1 of three standard clustering metrics for coreference resolution. Each such performance gap captures a different facet of gender bias, and one might be more interested in one of the metrics depending on the application.

We compute two types of performance gap metrics: (1) the sum of absolute gap values over all classes; (2) the Pearson correlation between the performance gap for a class and the percentage of women in that class. For instance, if y is a profession, we measure the correlation between performance gaps and percentages of women in each profession.[3] The two metrics are closely related but answer slightly different questions: the sum quantifies how a model behaves differently on different genders, and the correlation shows the relation of model behaviour to social biases (in the world or the data) without regard to actual gap size.

Statistical metrics. For breadth of analysis, we examine three additional statistical metrics (Barocas et al., 2019), which correspond to different notions of bias. All three are measured as differences (d) between two probability distributions, and we then obtain a single bias quantity per metric by summing all computed distances.

- *Independence*: $d\big(P(\mathbf{r}|\mathbf{z} = z), P(\mathbf{r})\big) \forall z \in \{F, M\}$. For instance, we measure the difference between the distribution of model's predictions on women and the distribution of all predictions. Independence is stronger as the prediction $\mathbf{r}$ is less correlated with the protected attribute $\mathbf{z}$. It is measured with no relation to the gold labels.
- *Separation*: $d\big(P(\mathbf{r}|\mathbf{y} = y, \mathbf{z} = z), P(\mathbf{r}|\mathbf{y} = y)\big)$ $\forall y \in \mathcal{Y}, z \in \{F, M\}$. For instance, we measure the difference between the distribution of a model's predictions on women who are teachers and the distribution of predictions on all teachers. It encapsulates the TPR and FPR gaps discussed previously, and can be seen as a more general metric.
- *Sufficiency*: $d\big(P(\mathbf{y}|\mathbf{r} = r, \mathbf{z} = z), P(\mathbf{y}|\mathbf{r} = r)\big)$. For instance, we measure the difference between the distribution of gold labels on women classified as teachers by the model and the distribution of gold labels on all individuals classified as teachers by the model. Sufficiency relates to the concept of calibration in classification. A difference in the classifier's scores for men and for women indicates that it might be penalizing or over-promoting one of the genders.

2.1.2 Intrinsic Metrics

Intrinsic methods are applied to the representation obtained from the feature extractor. These methods are independent of any downstream task. The intrinsic metric is a function:

$$I(g(\boldsymbol{X}), \boldsymbol{Z}) \in \mathbb{R}$$

Compression. Our main intrinsic metric is the *compression* of gender information evaluated by a minimum description length (MDL) probing classifier (Voita and Titov, 2020), trained to predict gender from the model's representations. Probing classifiers are widely used for predicting various properties of interest from frozen model representations (Belinkov and Glass, 2019). MDL probes were proposed because a probe's accuracy may be misleading due to memorization and other issues (Hewitt and Liang, 2019; Belinkov, 2021). We use the MDL online code, where the probe is trained in timesteps, on increasing subsets of the training set, then evaluated against the rest of it. Higher compression indicates greater gender extractability.

CEAT. We also measure CEAT (Guo and Caliskan, 2021), which is a contextualized version

[3] Percentages for coreference resolution are taken from labour statistics, following Zhao et al. (2018). For occupation prediction we use training set statistics following De-Arteaga et al. (2019), *before* balancing.

of WEAT (Caliskan et al., 2017), a widely used bias metric for static word embeddings. WEAT defines sets $\mathbb{X}$ and $\mathbb{Y}$ of target words, and sets $\mathbb{A}$ and $\mathbb{B}$ of attribute words. For instance, $\mathbb{A}$ and $\mathbb{B}$ contain males and females names, while $\mathbb{X}$ and $\mathbb{Y}$ contain career and family related words, respectively. The bias is operationalized as the geometric proximity between the target and attribute word embeddings, and is quantified in CEAT by the Combined Effect Size (CES) and a p-value for the null hypothesis of having no biased associations. For more information on CEAT refer to Appendix A.4.3.

2.2 Debiasing Techniques

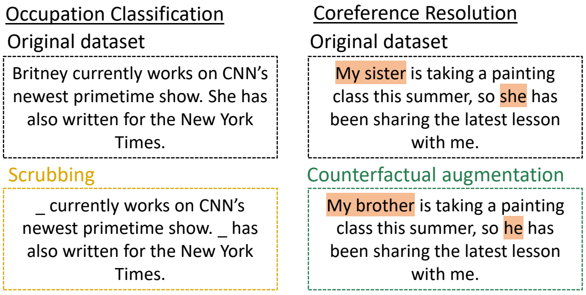
We debias models by modifying the downstream task's training data before fine-tuning. *Scrubbing* (De-Arteaga et al., 2019) removes first names and gender-specific terms ("he", "she", "husband", "wife", "Mr", "Mrs", etc.). *Balancing* subsamples or oversamples examples such that each gender is equally represented in the resulting dataset w.r.t each label. *Anonymization* (Zhao et al., 2018) removes named entities. *Counterfactual Augmentation* (Zhao et al., 2018) involves replacing male entities in an example with female entities, and adding the modified example to the training set. As some of these are dataset/task-specific, we give more details in the following section.

3 Experiments

In each experiment, we fine-tune a model for a downstream task. For training, we use either the original dataset or a dataset debiased with one of the methods from Section 2.2. Figure 2 presents examples of debiasing methods for the two downstream tasks. We measure two intrinsic metrics by probing that model's inner representations for gender extractability (as measured by MDL) and by CEAT, and test various extrinsic metrics. The relation between one intrinsic and one extrinsic metric becomes one data point, and we repeat over many random seeds (for both the model and the probe). Further implementation details are in appendix A.

3.1 Occupation Prediction

The task of occupation prediction is to predict a person's occupations (from a closed set), based on their biography. We use the Bias in Bios dataset (De-Arteaga et al., 2019). Regardless of the training method, the test set is subsampled such that each profession has equal gender representation.



Occupation Classification

Original dataset

Britney currently works on CNN's newest primetime show. She has also written for the New York Times.

Scrubbing

_ currently works on CNN's newest primetime show. _ has also written for the New York Times.

Coreference Resolution

Original dataset

My sister is taking a painting class this summer, so she has been sharing the latest lesson with me.

Counterfactual augmentation

My brother is taking a painting class this summer, so he has been sharing the latest lesson with me.

Figure 2: Examples of two debiasing methods performed on the data.

Model. Our main model is a RoBERTa model (Liu et al., 2019) topped with a linear classifier, which receives the [CLS] token embedding as input and generates a probability distribution over the professions. In addition, we experiment with training a baseline classifier layer on top of a frozen, non-finetuned RoBERTa. We also replicate our RoBERTa experiments with a DeBERTa model (He et al., 2020), to verify that our results are are not model specific and hold more broadly.

Debiasing Techniques. Following De-Arteaga et al. (2019) we experiment with scrubbing the training dataset. Figure 2 shows an example biography snippet and its scrubbed version. We also conduct balancing (per profession, subsampling and oversampling to ensure an equal number of males and females per profession), which has not previously been used on this dataset and task.

Metrics. We measure all bias metrics from Section 2.1 except for F1.

Probing. The probing dataset for this task is the test set, and the gender label of a single biography is the gender of the person described in it. We probe the [CLS] token representation of the biography. In addition to the models described above, we measure baseline extractability of gender information from a randomly initialized RoBERTa model.

3.2 Coreference Resolution

The task of coreference resolution is to find all textual expressions referring to the same real-world entities. We train on Ontonotes 5.0 (Weischedel et al., 2013) and test on the Winobias challenge dataset (Zhao et al., 2018). Winobias consists of sentence pairs, pro- and anti-stereotypical variants, with individuals referred to by their profession. For example, "The physician hired the secretary be-

			Extrinsic							
Debiasing Strategy	**Intrinsic**		**Before**				**After**			
	Compression	CEAT	TPR (P)	FPR (S)	Sep	Suff	TPR (P)	FPR (S)	Sep	Suff
Random	5.61*	0.12†	-	-	-	-	-	-	-	-
Pre-trained	10.12	0.49*	-	-	-	-	-	-	-	-
None	4.12	0.22	0.76	0.08	0.33	9.45	0.78	0.073	0.33	9.70
Oversampling	8.52*	0.29	0.73	0.09*	0.31	8.32*	0.81*	0.068*	0.33	10.91*
Subsampling	3.57	0.22	**0.32***	**0.03***	**0.20***	**1.22***	**0.70***	0.08*	0.30*	1.32*
Scrubbing	**1.70***	0.23	0.70*	0.06*	0.30	4.93*	0.71*	**0.06***	**2.56***	**0.81***

(a) Occupation classification: Comparison of intrinsic and extrinsic metrics before and after retraining of classification layer, over 10 seeds per fine-tuned model and per retrained classification model.

			Extrinsic							
Debiasing Strategy	**Intrinsic**		**Before**				**After**			
	Compression	CEAT	F1 diff	FPR (S)	Sep	Suff	F1 diff	FPR (S)	Sep	Suff
Random	0.83*	0.12†	-	-	-	-	-	-	-	-
Pre-trained	0.96	0.49*	-	-	-	-	-	-	-	-
None	1.98	0.35	6.63	0.12	1.25	8.69	6.07	0.11	1.19	7.35
Anon	2.07*	0.31*	7.26	0.13	1.34	8.82	7.42*	0.13*	1.34*	8.66*
CA	**1.50***	0.27*	**2.30***	0.05*	**0.54***	1.67*	3.67*	0.06*	0.67*	2.40*
Anon + CA	1.54*	**0.25***	2.42*	**0.049***	0.56*	**1.56***	**2.86***	**0.05***	**0.59***	**1.65***

(b) Coreference resolution: Comparison of intrinsic and extrinsic metrics before and after retraining of classification layer, over 10 seeds per fine-tuned model and 5 seeds per retrained classification model.

Table 1: Results on both tasks. * marks significant reduction or increase in bias ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy None). The lowest bias score in each column is marked with **bold**. P = Pearson; S = Sum. † was computed only on 3 out of 10 tests for which CEAT's $p < 0.05$.

cause *he/she* was busy." is pro/anti-stereotypical, based on US labor statistics. [4] A coreference system is measured by the performance gap between the pro- and anti-stereotypical subsets.

Model. We use the model presented in Lee et al. (2018a) with RoBERTa as a feature extractor.

Debiasing Techniques. Following Zhao et al. (2018), we apply anonymization (denoted as Anon) and counterfactual augmentation (CA) on the training set. These techniques were used jointly in previous work; we examine each individually as well.

Metrics. Following Zhao et al. (2018), we measure the F1 difference between anti- and pro-stereotypical examples.[5] We also interpret the task as a classification problem, and measure all metrics from Section 2.1. For more details refer to Appendix A.4.2.

Probing. We probe the representation of a profession word as extracted from Winobias sentences, after masking out the pronouns. We define a profession's gender as the stereotypical gender for this profession. To prevent memorization by the probe—given the small number of professions—the dataset is sorted so that professions are gradually added to the training set, so a success on the validation set is on previously unseen professions.

4 Results

Tables 1a and 1b present intrinsic and extrinsic metrics for RoBERTa models on the occupation prediction and coreference resolution tasks, respectively. We present a representative subset of the measured metrics that demonstrate the observed phenomena; full results are found in Appendix B. The DeBERTa model results are consistent with the RoBERTa model trends.

4.1 Compression Reflects Debiasing Effects

As shown in the tables, compression captures differences in models that were debiased differently. CEAT, however, cannot differentiate between occupation prediction models. For example, in occupation prediction (Table 1a) the compression rate

[4] Labor Force Statistics from the Current Population Survey, https://www.bls.gov/cps/cpsaat11.htm

[5] We combined the T1 and T2 datasets, as well as the dev and test datasets, to create a single held-out challenge dataset.

varies significantly between a non-debiased and a debiased model via scrubbing and oversampling, while CEAT detects no difference between the models. In coreference resolution (Table 1b), both compression and CEAT are able to identify differences between the non-debiased model and the others, such as CA, which has both a lower compression and CEAT effect. But the CEAT effect sizes are small (below 0.5), which implies no bias, in contrast to the extrinsic metrics.

4.2 High Gender Extractability Implies Superficial Debiasing

Extrinsic and intrinsic effects of debiasing. In occupation classification (Table 1a), somewhat surprisingly, subsampling the training data has the strongest effect on extrinsic metrics, but not on compression rate. Scrubbing reduces both intrinsic and extrinsic metrics, although its effect on extrinsic metrics is limited compared to subsampling. Training with oversampling caused less reduction in extrinsic bias metrics. A consequence of oversampling is that some metrics are less biased, but compression rates are increased, so gender information is more accessible. The effectiveness of subsampling over other metrics is further discussed in appendix C. In coreference resolution (Table 1b), while both CA and CA with anonymization reduced gender extractability as well as external bias metrics, anonymization alone *increased* intrinsic bias without affecting external bias metrics significantly.

Debiasing without fine-tuning. As the effect on extrinsic bias did not match the effect on intrinsic bias in several cases, we examined the role of the classification layer. We trained a model for occupation prediction without fine-tuning the underlying RoBERTa model. Training on a subsampled dataset also reduced the extrinsic metrics (0.15, 0.03, 0.20, and 0.31, respectively, on TPR gaps Pearson, FPR gaps sum, separation sum, and sufficiency sum). Detailed results of this experiment can be found in Appendix B. Since no updates were made to the LM, the internal representations could not be debiased, thus the debiasing observed in this model can only be superficial.

Retraining the classification layer. Fine-tuning of both tasks revealed that lower extrinsic metrics did not always lead to lower compression. Does this indicate cases where the debiasing process is only superficial, and the internal representations remain biased? To test this hypothesis, we froze the previously fine-tuned LM's weights, and retrained the classification layer. We used the original (non-debiased) training set for retraining.

Tables 1a and 1b also compare extrinsic metrics before and after retraining. All models show bias restoration, due to the classification layer being trained on the biased dataset.[6] The amount of bias restored varies between models in a way that is predictable by the compression metric.

In the occupation prediction task, comparing Before and After numbers in Table 1a, the model fine-tuned using a scrubbed dataset—which has the lowest compression rate—displays the least bias restoration, confirming that the LM absorbed the process of debiasing. The model fine-tuned on subsampled data has higher extrinsic bias after retraining. Hence, the debiasing was primarily cosmetic, and the representations within the LM were not debiased. The model fine-tuned on oversampled data—which has the highest compression—has the highest extrinsic bias (except for FPR), even though this was not true before retraining.

In coreference resolution, comparing Before and After numbers in Table 1b, models with the least extrinsic bias (CA and CA+Anon) are also least biased after retraining. Compression rate predicted this; these models also had lower compression rates than non-debiased models. Interestingly, the model fine-tuned with an anonymized dataset is the most biased after retraining, consistent with its high compression rate relative to the other models. As with subsampling and oversampling in occupation prediction, anonymization's (lack of) effect on extrinsic metrics was cosmetic (compare None and Anon in Before block, Table 1b). Anonymization actually had a biasing effect on the LM, which was realized after retraining.

We conclude that compression rate is a useful indicator of superficial debiasing, and can potentially be used to verify and gain confidence in attempts to debias an NLP model, especially when there is little or no testing data.

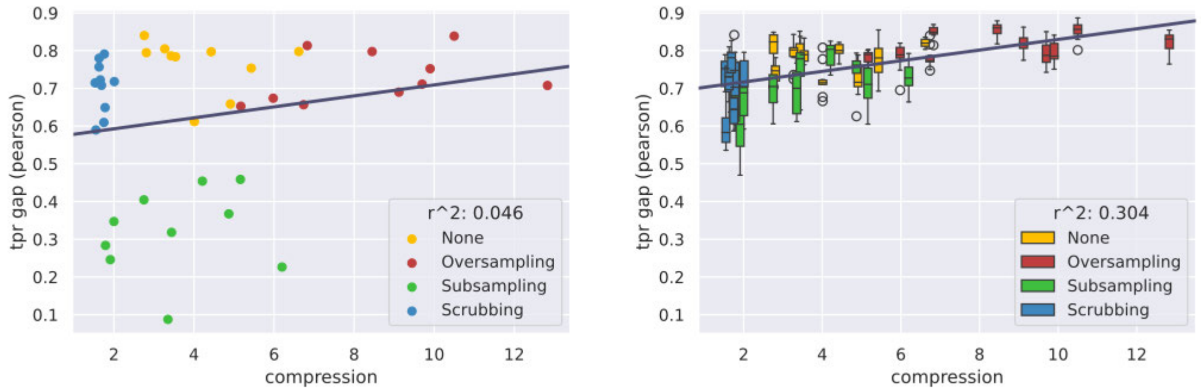
4.3 Correlation between Extrinsic and Intrinsic Metrics

Table 2 shows correlations between compression rate and various extrinsic metrics before and after

[6] The training datasets contain bias. The occupation prediction set has an unbalanced amount of males and females per profession (for example 15% of software engineers are females). The coreference resolution training set has more male than female pronouns, and males are more likely to be referred to by their profession (Zhao et al., 2018).

	Occupation Classification				Coreference Resolution			
	R^2 Compression		R^2 CEAT		R^2 Compression		R^2 CEAT	
Metric	Before	After	Before	After	Before	After	Before	After
F1 diff ($pro - anti$)	-	-	-	-	0.821	0.709	0.246	0.005
TPR gap (P)	0.046	0.304	0.042	0.049	0.222	0.006	0.008	0.012
TPR gap (S)	0.049	0.449	0.022	0.036	0.817	0.752	0.297	0.003
FPR gap (P)	0.001	0.120	0.008	0.002	0.021	0.054	0.002	0.000
FPR gap (S)	0.353	0.046	0.079	0.001	0.844	0.773	0.263	0.004
Precision gap (P)	0.032	0.173	0.000	0.000	0.068	0.038	0.019	0.000
Precision gap (S)	0.174	0.529	0.000	0.021	0.849	0.774	0.268	0.006
Independence gap (S)	0.251	0.382	0.050	0.005	0.778	0.732	0.355	0.001
Separation gap (S)	0.066	0.165	0.046	0.009	0.835	0.776	0.261	0.005
Sufficiency gap (S)	0.202	0.567	0.040	0.034	0.825	0.753	0.287	0.002

Table 2: Coefficient determination of the regression line taken on the compression rate or CEAT and each extrinsic metric, before and after retraining of the classification layer. P = Pearson; S = Sum.



(a) Fine-tuned models. Each point is a single seed for training and testing the model.

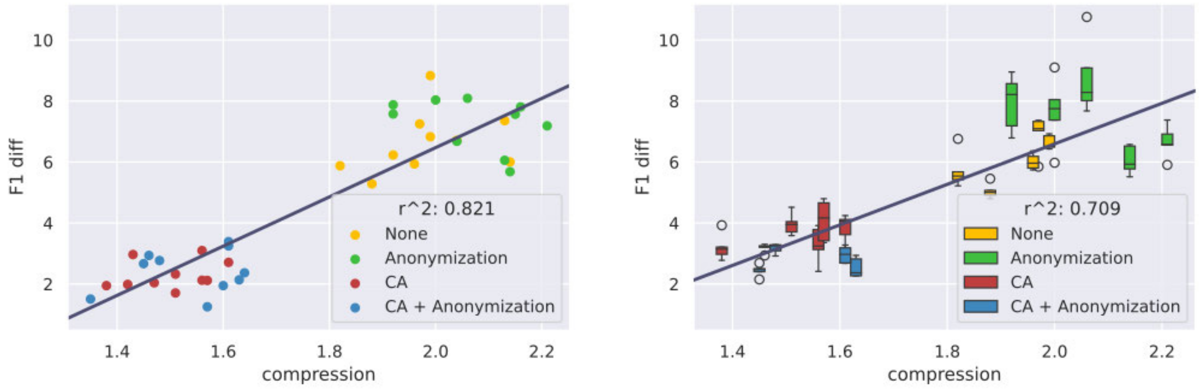
(b) After retraining. Each box represents 10 runs of retraining on the same fine-tuned feature extractor.

Figure 3: Occupation prediction: Compression vs. TPR-gap (Pearson) after various debiasing strategies.

retraining. In occupation prediction, certain extrinsic metrics have a weak correlation with compression rate, while others do not. Except one metric (FPR gap sum), the compression rate and the extrinsic metric correlate more after retraining. Figure 3 illustrates this for TPR-gap (Pearson). The increase is due to superficial debiasing, especially by subsampling data, which prior to retraining had low extrinsic metrics and relatively high intrinsic metrics. This shows that correlation between extrinsic metrics and compression rate for certain metrics is stronger than it appeared before retraining. It is unsurprising that CEAT does not correlate with any extrinsic metrics, since CEAT could not distinguish between different models.

Coreference resolution shows stronger correlations between compression rate and extrinsic metrics, but low correlations between Pearson metrics. We further discuss cases of no correlation in appendix D. Correlations decrease after retraining, but metrics that were highly correlated remain so ($>$ 0.7 after retraining). The correlations are visualized for F1 difference metrics in Figure 4. CEAT and extrinsic metrics correlate much less than compression rate (Table 2). Our results are in line with those of Goldfarb-Tarrant et al. (2021), who found a lack of correlation between extrinsic metrics and WEAT, the static-embedded version of CEAT.

Given that recent work (Goldfarb-Tarrant et al., 2021; Cao et al., 2022) questions the validity of intrinsic metrics as a reliable indicator for gender bias, the compression rate provides a reliable alternative to current intrinsic metrics, by offering correlation to many extrinsic bias metrics.



(a) Fine-tuned models. Each point is a single seed for training and testing the model.

(b) After retraining. Each box represents 5 runs of retraining on the same fine-tuned feature extractor.

Figure 4: Coreference resolution: Compression vs. F1 difference after various debiasing strategies.

5 Related Work

There are few studies that examine both intrinsic and extrinsic metrics. Previous work by Goldfarb-Tarrant et al. (2021) showed that debiasing static embeddings intrinsically is not correlated with extrinsic bias, challenging the assumption that intrinsic metrics are predictive of bias. We examine the other direction, exploring how extrinsic debiasing affects intrinsic metrics. We also extend beyond their work to contextualized embeddings, a wider range of extrinsic metrics, and a new, more effective intrinsic metric based on information-theoretic probing. A contemporary work by Cao et al. (2022) measured the correlations between intrinsic and extrinsic metrics in contextualized settings across different language models. In contrast, our work examines the correlations across different versions of the same language model by fine-tuning it using various debiasing techniques.

Studies that inspect extrinsic metrics include either a challenge dataset curated to expose differences in model behavior by gender, or a test dataset labelled by gender. Among these datasets are Winobias (Zhao et al., 2018), Winogender (Rudinger et al., 2018) and GAP (Webster et al., 2018) for coreference resolution, WinoMT (Stanovsky et al., 2019) for machine translation, EEC (Kiritchenko and Mohammad, 2018) for sentiment analysis, BOLD (Dhamala et al., 2021) for language generation, gendered NLI (Sharma et al., 2020) for natural language inference and Bias in Bios (De-Arteaga et al., 2019) for occupation prediction.

Studies that measure gender bias intrinsically in static word or sentence embeddings measure characteristics of the geometry, such as the proximity between female- and male-related words to stereotypical words, or how embeddings cluster or relate to a gender subspace (Bolukbasi et al., 2016; Caliskan et al., 2017; Gonen and Goldberg, 2019b; Ethayarajh et al., 2019). However, metrics and debiasing methods for static embeddings do not apply directly to contextualized ones. Several studies use sentence templates to adapt to contextual embeddings (May et al., 2019; Kurita et al., 2019; Tan and Celis, 2019). This templated approach is difficult to scale, and lacks the range of representations that a contextual embedding offers. Other work extracts embedding representations of words from natural corpora (Zhao et al., 2019; Guo and Caliskan, 2021; Basta et al., 2019). These studies often adapt the WEAT method (Caliskan et al., 2017), which measures embedding geometry. None measure the effect of the presumably found "bias" on a downstream task.

There is a growing conversation in the field (Barocas et al., 2017; Kate Crawford, 2017; Blodgett et al., 2020; Bommasani et al., 2021) about the importance of articulating the harms of measured bias. In general, extrinsic metrics have clear, interpretable impacts for which harm can be defined. Intrinsic metrics have an unclear effect. Without evidence from a concrete downstream task, a found intrinsic bias is only theoretically harmful. Our work is a step towards understanding whether intrinsic metrics provide valuable insights about bias in a model.

6 Discussion and Conclusions

This study examined whether bias in internal representations is related to extrinsic bias. We designed a new framework in which we debias a model on a downstream task, and measure its intrinsic bias. We found that gender extractability from internal representations, measured by compression rate via MDL probing, reflects bias in a model. Compression was much more reliable than an alternative intrinsic metric for contextualised representations, CEAT. Compression correlated well—to varying degrees—with many extrinsic metrics. We thus encourage NLP practitioners to use compression as an intrinsic indicator for gender bias in NLP models. When comparing two alternative models, a lower compression rate provides confidence in a model's superiority in terms of gender bias. The relative success of compression over CEAT may be because the compression rate was calculated on the same dataset as the extrinsic metrics, whereas CEAT was measured on a different dataset not necessarily aligned with a specific downstream task. The use of a non-task-aligned dataset is a common strategy among other intrinsic metrics (May et al., 2019; Kurita et al., 2019; Basta et al., 2021). Another possible explanation is that compression rate measures a more focused concept, namely the gender information within the internal representations. CEAT measures proximity among embeddings of general terms that may include other social contexts that do not directly relate to gender (e.g. a female term like 'lady' or 'Sarah' contains information about not just gender but class, culture, formality, etc, and it can be hard to isolate just one of these from the rest).

Our results show that when a debiasing method reduces extrinsic metrics but not compression, it indicates that the language model remains biased. When such superficial debiasing occurs, the debiased language model may be reapplied to another task, as in Jin et al. (2021), resulting in unexpected biases and nullifying the supposed debiasing. Our findings suggest that practitioners of NLP should take special care when adopting previously debiased models and inspect them carefully, perhaps using our framework. Our results differ from those of Mendelson and Belinkov (2021a), who found that the debiasing increases bias extractability as measured by compression rate. However, they studied different, non-social biases, that arise from spurious or unintended correlations in training datasets (often called dataset biases). In our case, some debiasing strategies increase intrinsic bias while others decrease it. Future work could investigate why debiasing affects extractability differently for these two types of biases.

Our work also highlighted the importance of the classification layer. Using a debiased objective, such as a balanced dataset, the classification layer can provide significant debiasing. This holds even if the internal representations are biased and the classifier is a single linear layer, as shown in the occupation prediction task. Bias stems in part from internal LM bias and in part from classification bias. Practitioners should focus their efforts on both parts when attempting to debias a model.

We used a broader set of extrinsic metrics than is typically used, and found that the bias metrics behaved differently: some decreased more than others after debiasing, and they correlated differently with compression rate. Debiasing efforts may not be fully understood by testing only a few extrinsic metrics. However, compression as an intrinsic bias metric can indicate meaningful debiasing of internal model representations even when not all metrics are easily measurable, since it correlates well with many extrinsic metrics.

A major limitation of this study is the use of gender as a binary variable, which is trans-exclusive. Cao and Daumé III (2020) made the first steps towards inclusive gender bias evaluation in NLP, revealing that coreference systems fail on gender-inclusive text. Further work is required to adjust our framework to non-binary genders, potentially revealing insights about the poor performance of NLP systems in that area.

Acknowledgements

This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 448/20) and by an Azrieli Foundation Early Career Faculty Fellowship. We also thank Kate McCurdy and Andreas Grivas for comments on early drafts, the members of the Technion CS NLP group for their valuable feedback, and the anonymous reviewers for their useful suggestions.

References

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

Solon Barocas, Moritz Hardt, and Arvind Narayanan.

2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2021. Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications*, 33(8):3371–3384.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *Computational Linguistics 2021*.

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Su Lin Blodgett, Solon Barocas, Hal Daum'e, and H. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *ACL*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Aylin Caliskan, J. Bryson, and A. Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183 – 186.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. *arXiv preprint arXiv:2203.13928*.

Maria De-Arteaga, Alexey Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. C. Geyik, K. Kenthapadi, and A. Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

J. Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Y. Goldberg. 2019a. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NAACL-HLT*.

Hila Gonen and Yoav Goldberg. 2019b. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online. Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias. keynote at neurips.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Kweku Kwegyir-Aggrey, Rebecca Santorella, and Sarah M. Brown. 2021. Everything is relative: Understanding fairness with optimal transport. *ArXiv*, abs/2102.10349.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018a. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018b. Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Mendelson and Yonatan Belinkov. 2021a. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1557, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Mendelson and Yonatan Belinkov. 2021b. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Marnie E Rice and Grant T Harris. 2005. Comparing effect sizes in follow-up studies: Roc area, cohen's d, and r. *Law and human behavior*, 29(5):615–620.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a name? reducing bias in bios without access to protected attributes. In *Proceedings of NAACL-HLT*, pages 4187–4195.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2020. Evaluating gender bias in natural language inference. In *NeurIPS 2020 Workshop on Dataset Curation and Security*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on wikipedia.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA. Association for Computing Machinery.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5:1–24.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

R. Weischedel, E. Hovy, M. Marcus, and Martha Palmer. 2013. Ontonotes : A large training corpus for enhanced processing. *LDC2013T19, Philadelphia, Penn.: Linguistic Data Consortium.*

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Implementation Details

We used RoBERTa in all models (base size, 120M parameters). We use following random seeds in all repeated experiments: 0, 5, 11, 26, 42, 46, 50, 63, 83, 90. Our code was implemented mainly using the Python libraries Pytorch (Paszke et al., 2019), Transformers (Wolf et al., 2020), Sklearn (Pedregosa et al., 2011), and the experiments were logged using Wandb (Biewald, 2020).

A.1 Occupation Classification

We fine-tuned a RoBERTa-base model with a linear classification layer on top. Training was done for 10 epochs at a learning rate of 5e-5, batch size of 64. The input to RoBERTa was the biography tokens, which is limited to the first 128 tokens. The resulting [CLS] token embedding is fed to the classifier to predict the occupation. The probing task involves using the same [CLS] token and training the probing classifier to predict the gender of the person in the biography. The experiments without fine-tuning included either a pre-trained or a previously fine-tuned RoBERTa. We first extracted the pre-trained RoBERTa's embeddings of tokens from the [CLS] and then trained a linear classifier on them. The learning rate was 0.001 and the batch size was 64. We trained the classification layer with pre-trained RoBERTa on 300 epochs, but with fine-tuned RoBERTa, 10 epochs were sufficient. For all training processes, the epoch with the greatest validation accuracy was saved. Fine-tuning took 7 hours on a GeForce RTX 2080 Ti GPU. Bias in Bios contains almost 400k biographies, and we obtain validation (10%) and test set (25%) by splitting with Scikit-learn's (Pedregosa et al., 2011) test_train_split with our random seeds.

A.2 Coreference Resolution

We use the implementation of Xu and Choi (2020), a model that was introduced by Lee et al. (2018b) and has been adopted by many coreference resolution models. Coreference resolution is the process of clustering different mentions in a text that refer to the same real-world entities. The task is solved by detecting mentions through text spans and then predicting for each pair of spans if they represent the same entity. The span representations were extracted with a RoBERTa model, which is fine-tuned throughout the training process, except in the retraining experiment. Fine-tuning took 3 hours on an NVIDIA RTX A6000 GPU. Ontonotes 5.0 has 625k sentences and we use the standard validation and test splits.

A.3 Probing Classifier

We use the MDL probe (Voita and Titov, 2020) implementation by Mendelson and Belinkov (2021b). In all experiments, we use a linear probe and train it with a batch size of 16 and a learning rate of 1e-3. The timestamps used, meaning the accumulating fractions of data that the probe is trained on, are 2.0%, 3.0%, 4.4%, 6.5%, 9.5%, 14.0%, 21.0%, 31.0%, 45.7%, 67.6%, 100%.

A.4 Metrics

A.4.1 Fairness-Based Metrics Implementation

All three statistical fairness metrics measure the difference between two probability distributions, where this difference describes a notion of bias. We calculate Independence and Separation via Kullback–Leibler (KL) divergence, using the AllenNLP implementation (`https://github.com/allenai/allennlp`). We calculate Sufficiency via Wasserstein distance instead, which is motivated by Kwegyir-Aggrey et al. (2021). In this case, we cannot use KL divergence, since there are some classes that do not occur in model predictions for both male and female genders. This causes the probability distributions to not have the same support, and KL divergence is unbounded. Wasserstein distance lacks the requirement for equal support.

A.4.2 Classification Metrics Interpretation in Winobias

Winobias datasets contain pairs of stereotypical and anti-stereotypical sentences. The stereotypes are derived from the US labor statistics (for instance, a profession with a majority of males is stereotypically male). Since coreference resolution is viewed as a clustering problem, it is usually measured via clustering evaluation metrics. Coreference resolution is commonly measured as the average F1 score of these, and the same is true for Winobias. Nevertheless, coreference resolution is accomplished by making a prediction for each pair of mentions, so it can be seen as a classification task. Winobias can be viewed as a simpler task than general coreference resolution, as it contains exactly two mentions of professions and one pronoun, which refers to exactly one profession. Therefore, we reframe it as a classification problem. In a Winobias sentence with two professions x and y, as well as a pronoun p, where p is referring to x, a true positive would be to cluster x and p together, while a false positive would be to cluster y and p together. Our classification metrics are derived based on these definitions. For instance, the TPR gap for profession “teacher”, which is a stereotypical female occupation, is the TPR rate on pro-stereotypical sentences (with a female pronoun) minus the TPR rate on anti-stereotypical sentences (with a male pronoun).

A.4.3 CEAT

The Word Embedding Association Test (WEAT) developed by (Caliskan et al., 2017) is a method for evaluating bias in static word embeddings. The test is defined as follows: given two sets of target words $\mathbb{X}$, $\mathbb{Y}$ (e.g., 'executive', 'management', 'professional' and 'home', 'parents', 'children') and two sets of attribute words (e.g., male names and female names), and using $\vec{w}$ to represent the word embedding for word w, the effect size is:

$$\text{ES} = \text{mean}_{x\in\mathbb{X}} s(x,\mathbb{A},\mathbb{B}) - \text{mean}_{y\in\mathbb{Y}} s(y,\mathbb{A},\mathbb{B})$$

where

$$s(x,\mathbb{A},\mathbb{B}) = \frac{\text{mean}_{a\in\mathbb{A}}\cos(\vec{\boldsymbol{x}},\vec{\boldsymbol{a}}) - \text{mean}_{a\in\mathbb{A}}\cos(\vec{\boldsymbol{x}},\vec{\boldsymbol{b}})}{\text{std-dev}_{w\in\mathbb{X}\bigcup\mathbb{Y}} s(w,\mathbb{A},\mathbb{B})}$$

In essence, the effect size measures how different are the distances between the embedding vectors of each target group and the attribute groups. Specifically, if $s(x, \mathbb{A}, \mathbb{B}) > 0$, $\vec{\boldsymbol{x}}$ is more similar to attribute words $\mathbb{B}$ and vice versa. For instance, a larger effect size is observed if target words $\mathbb{X}$ are more similar to attribute words $\mathbb{A}$ and target words $\mathbb{Y}$ are more similar to attribute words $\mathbb{B}$. $|ES| > 0.5$ and $|ES| > 0.8$ are considered medium and large effect sizes, respectively (Rice and Harris, 2005). The null hypothesis holds that there is no difference between the two sets of target words in terms of their relative similarity to the two sets of attribute words, indicating that there are no biased associations. Statistical significance is defined by the p-value of WEAT, which reflects the probability of observing the effect size under the null hypothesis.

Since a word can take on a great variety of vector representations in a contextual setting, ES varies according to the sentences used to extract word representation. Thus, to adopt WEAT to contextualized representations, the Combined Effect Size (CES) (Guo and Caliskan, 2021) is derived as the distribution of WEAT effect sizes over many possible contextual word representations:

$$\text{CES}(\mathbb{X},\mathbb{Y},\mathbb{A},\mathbb{B}) = \frac{\sum_{i=1}^{N} v_i ES_i}{\sum_{i=1}^{N} v_i}$$

where ES_i denotes the WEAT effect size of the i'th choice of word representations from a large corpus, and v_i is the inverse of the sum of in-sample variance V_i and between-sample variance in the distribution of random-effects. As in Guo and Caliskan (2021), the representation for each word is derived from 10,000 random sentences extracted from a corpus of Reddit comments.

The combined effect size of each of the models is examined on WEAT stimulus 6, which contains target words of career/family and attribute words of male/female names. This was the only one that detected bias on a pre-trained RoBERTa (CES close to 0.5 and $p < 0.05$). The points that we kept in our analysis are those where $p < 0.05$, which make up 90% of the points in occupation prediction and 95% of the points in coreference resolution.

B Full Results

In this section we provide the full results of a RoBERTa model trained on the downstream task.

Table 3 presents results for the occupation prediction task after fine-tuning, Table 4 presents the retrained model results.

Figure 5 illustrates the correlations between extrinsic metrics and compression rate before and after retraining.

Table 5 presents the complete results for the occupation prediction task of the model trained without fine-tuning, meaning that the RoBERTa model is the pretrained version from Liu et al. (2019) and only the classification layer was updated. Subsampling the dataset has significant debiasing effects, which suggests that this debiasing method can achieve low extrinsic bias even when internal bias exists. The Pearson correlation on precision exhibits a different behavior. It makes sense nonetheless: precision is computed as $TP \backslash (TP + FP)$. A biased model will assign more examples of a specific profession to a specific gender (which aligns with the percentage of biographies of this profession with this gender on the training set), increasing both TP and FP and decreasing precision. The results on the coreference resolution task align with the results of occupation prediction.

Table 6 presents the results using a DeBERTa model (He et al., 2020) for the occupation classification task. The trends are similar to those of RoBERTa, with the same metrics showing an increase, no change, or decrease in correlation after re-training, suggesting a general trend in the behavior of these metrics in relation to internal model representations.

Table 7 displays the results on a finetuned model for the coreference resolution task and Table 8 displays the retraining results.

Figure 6 shows the correlations between compression rate and extrinsic metrics before and after the retraining.

	Debiasing Strategy			
Metric	None	Oversampling	Subsampling	Scrubbing
Compression	4.121 ± 1.238	8.522* ± 2.354	3.568 ± 1.516	**1.699*** ± 0.138
Accuracy	**0.861** ± 0.005	0.852* ± 0.004	**0.861** ± 0.003	0.851* ± 0.003
TPR gap (P)	0.763 ± 0.071	0.729 ± 0.067	**0.319*** ± 0.114	0.704* ± 0.068
TPR gap (S)	2.391 ± 0.257	2.145* ± 0.220	**1.598*** ± 0.273	2.019* ± 0.262
FPR gap (P)	0.591 ± 0.052	0.491* ± 0.059	**0.087*** ± 0.094	0.552 ± 0.063
FPR gap (S)	0.075 ± 0.010	0.085* ± 0.011	**0.030*** ± 0.006	0.057* ± 0.007
Precision gap (P)	-0.880 ± 0.031	-0.855 ± 0.115	**-0.299*** ± 0.215	-0.815* ± 0.040
Precision gap (S)	3.621 ± 0.337	3.401 ± 0.667	**1.549*** ± 0.229	2.590* ± 0.279
Independence gap (S)	0.009 ± 0.002	0.008 ± 0.002	**0.001*** ± 0.000	0.005* ± 0.001
Separation gap (S)	0.327 ± 0.051	0.305 ± 0.030	**0.204*** ± 0.032	0.296 ± 0.053
Sufficiency gap (S)	9.451 ± 1.945	8.324* ± 1.537	**1.218*** ± 0.330	4.930* ± 0.927

Table 3: Occupation Prediction: Results on a RoBERTa-based model trained over 10 seeds. Significant reduction or increase in a metric ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy is None), is marked with *. The lowest bias score or highest performance metric in each column is marked with **bold**. P = Pearson; S = Sum.

	Debiasing Strategy			
Metric	None	Oversampling	Subsampling	Scrubbing
Compression	4.121 ± 1.238	8.522 ± 2.354	3.568 ± 1.516	1.699 ± 0.138
Accuracy	0.859 ± 0.004	0.856 ± 0.003	0.853 ± 0.003	0.854 ± 0.003
TPR gap (P)	0.777 ± 0.047	0.813* ± 0.040	**0.704*** ± 0.075	0.714* ± 0.068
TPR gap (S)	2.482 ± 0.238	2.593* ± 0.240	2.164* ± 0.284	**1.989*** ± 0.227
FPR gap (P)	0.596 ± 0.041	0.603 ± 0.047	0.602 ± 0.041	**0.536*** ± 0.038
FPR gap (S)	0.073 ± 0.008	0.068* ± 0.007	0.081* ± 0.012	**0.059*** ± 0.005
Precision gap (P)	-0.877 ± 0.027	-0.891* ± 0.023	-0.889* ± 0.035	**-0.817*** ± 0.058
Precision gap (S)	3.710 ± 0.251	3.996* ± 0.272	3.555* ± 0.598	**2.703*** ± 0.255
Independence gap (S)	0.009 ± 0.002	0.010* ± 0.002	0.009 ± 0.003	**0.005*** ± 0.001
Separation gap (S)	0.334 ± 0.050	0.328 ± 0.048	0.300* ± 0.049	**0.274*** ± 0.041
Sufficiency gap (S)	9.701 ± 1.305	10.908* ± 1.354	8.370* ± 2.558	**5.239*** ± 0.798

Table 4: Occupation Prediction after retraining: Results on a RoBERTa-based model after retraining of the classification layer with 10 seeds for each pre-trained model. Significant reduction or increase in a metric ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy is None), is marked with *. The lowest bias score or highest performance metric in each column is marked with **bold**. P = Pearson; S = Sum.

	Debiasing Strategy			
Metric	None	Oversampling	Subsampling	Scrubbing
Accuracy	0.824 ± 0.003	0.815* ± 0.005	**0.831*** ± 0.001	0.807* ± 0.003
TPR gap (P)	0.839 ± 0.011	0.443* ± 0.053	**0.158*** ± 0.156	0.814 ± 0.029
TPR gap (S)	3.088 ± 0.192	**1.545*** ± 0.177	1.621* ± 0.088	3.154 ± 0.332
FPR gap (P)	0.598 ± 0.016	0.369* ± 0.029	**0.067*** ± 0.050	0.550* ± 0.012
FPR gap (S)	0.087 ± 0.004	0.041* ± 0.004	**0.027*** ± 0.003	0.112* ± 0.005
Precision gap (P)	-0.872 ± 0.028	-0.427* ± 0.074	**-0.161*** ± 0.162	-0.853 ± 0.019
Precision gap (S)	3.811 ± 0.253	1.736* ± 0.108	**1.551*** ± 0.195	3.907 ± 0.184
Independence gap (S)	0.014* ± 0.002	0.001* ± 0.000	**0.000*** ± 0.000	0.022* ± 0.001
Separation gap (S)	0.336* ± 0.044	0.214* ± 0.038	**0.203*** ± 0.024	0.432* ± 0.048
Sufficiency gap (S)	12.019* ± 1.721	2.105* ± 0.576	**1.478*** ± 0.394	13.798* ± 0.966

Table 5: Occupation Prediction: Results on a RoBERTa-based model trained without fine-tuning, over 5 seeds. The compression rate computed on a pre-trained RoBERTa model is 10.122. Significant reduction or increase in a metric ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy is None), is marked with *. The lowest bias score or highest performance metric in each column is marked with **bold**. P = Pearson; S = Sum.

	R^2 Compression		R^2 CEAT	
Metric	Before	After	Before	After
TPR gap (P)	0.023	0.120	0.051	0.006
TPR gap (S)	0.000	0.200	0.036	0.098
FPR gap (P)	0.083	0.153	0.121	0.149
FPR gap (S)	0.055	0.013	0.009	0.021
Precision gap (P)	0.002	0.135	0.046	0.031
Precision gap (S)	0.024	0.362	0.026	0.103
Independence gap (S)	0.034	0.084	0.0	0.054
Separation gap (S)	0.000	0.117	0.008	0.009
Sufficiency gap (S)	0.016	0.250	0.046	0.042

Table 6: Results for a DeBERTa model trained on occupation prediction task. Coefficient determination of the regression line taken on the compression rate or CEAT and each extrinsic metric, before and after retraining of the classification layer. P = Pearson; S = Sum. Coefficients are of lower magnitude for DeBERTa than RoBERTa models, but the same trends apply. They largely increase after retraining (save for FPR gap, and a few of the very low magnitude Pearson metrics). The increase after retraining does not apply to CEAT, and the correlations with CEAT are usually lower.

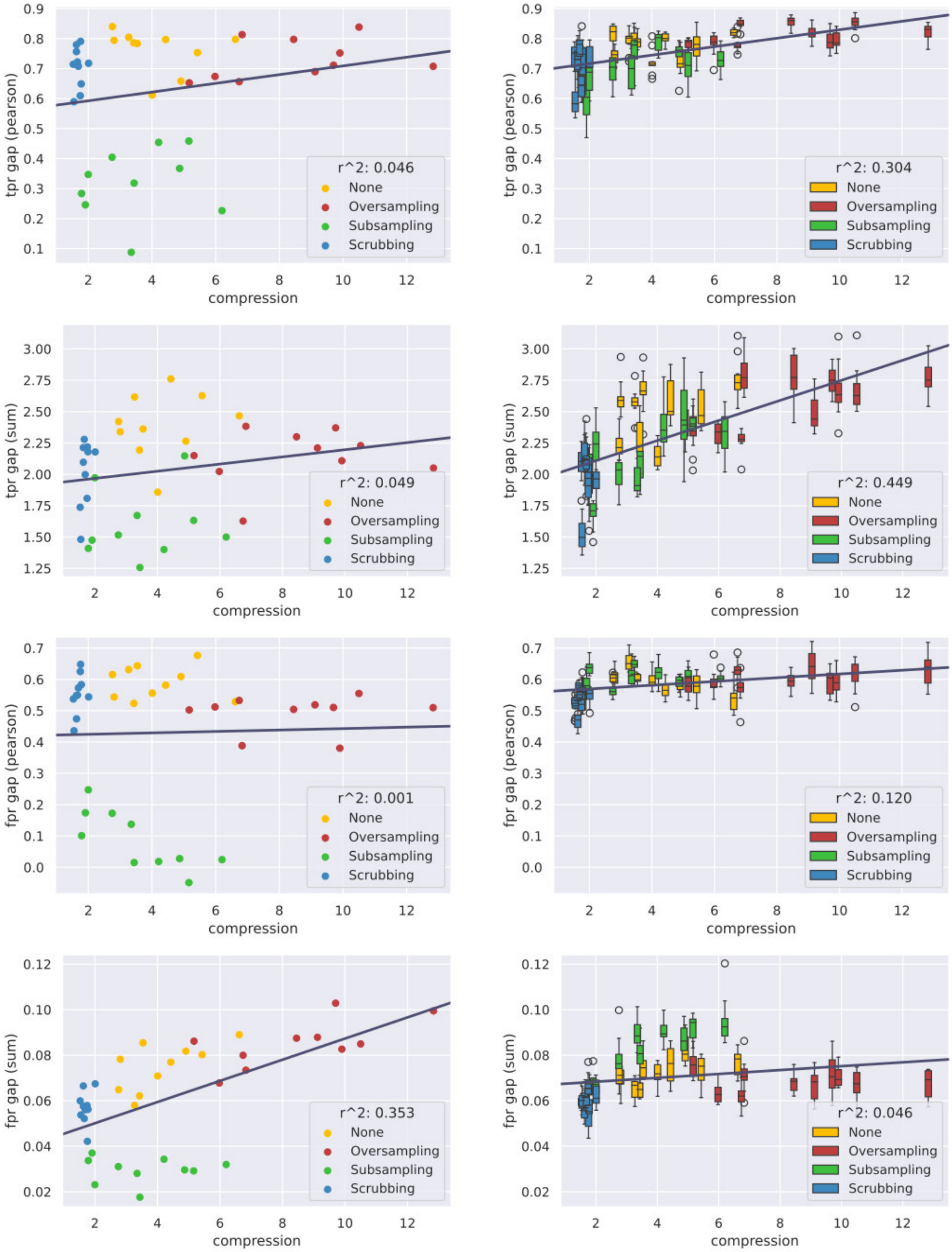


Figure 5: Occupation prediction: Before (left) and after (right) plots of compression rate versus and extrinsic metric. Cases of low correlation are discussed in D.1.

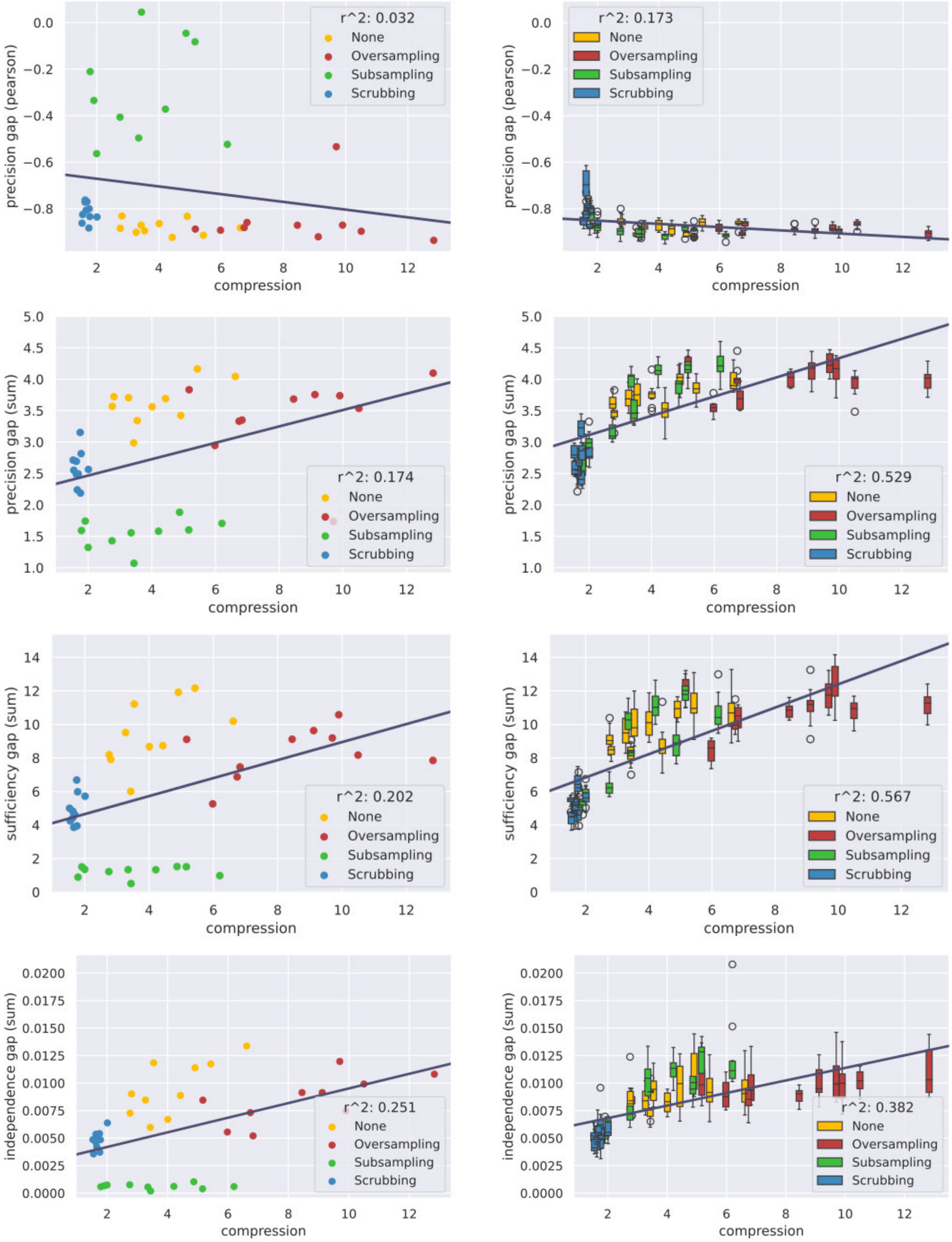


Figure 5: Occupation prediction: Before (left) and after (right) plots of compression rate versus and extrinsic metric. Cases of low correlation are discussed in D.1.

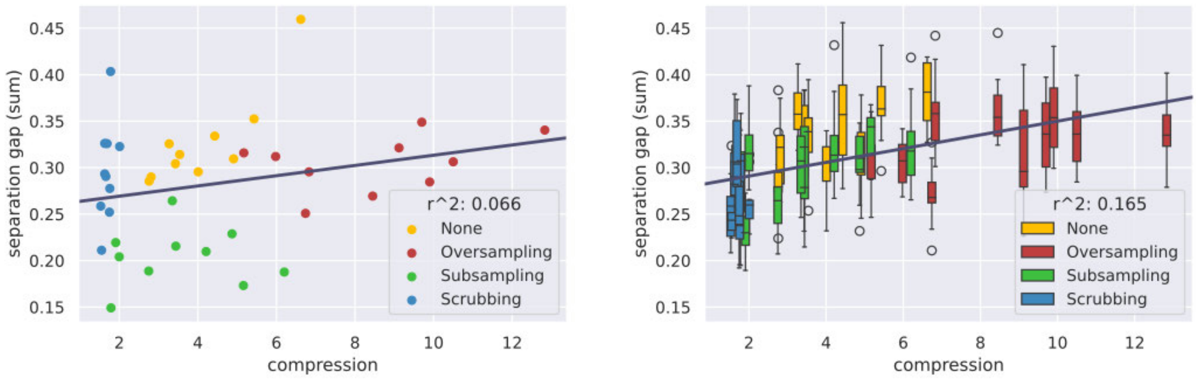


Figure 5: Occupation prediction: Before (left) and after (right) plots of compression rate versus and extrinsic metric. Cases of low correlation are discussed in D.1.

	Debiasing Strategy			
Metric	None	Anon	CA	Anon + CA
Compression	1.984 ± 0.101	2.073* ± 0.102	**1.502*** ± 0.075	1.540* ± 0.098
F1 (Ontonotes test)	76.406 ± 0.165	76.538 ± 0.176	77.187* ± 0.071	**77.246*** ± 0.230
F1 diff ($pro - anti$)	6.631 ± 1.013	7.256 ± 0.846	**2.302*** ± 0.466	2.422* ± 0.714
TPR gap (P)	0.654 ± 0.069	0.710* ± 0.047	**0.607** ± 0.082	0.627 ± 0.100
TPR gap (S)	4.884 ± 0.698	4.870 ± 0.509	2.041* ± 0.228	**2.014*** ± 0.286
FPR gap (P)	0.602 ± 0.036	0.620 ± 0.056	**0.572** ± 0.078	0.629 ± 0.107
FPR gap (S)	0.120 ± 0.015	0.128 ± 0.011	0.050* ± 0.006	**0.049*** ± 0.007
Precision gap (P)	-0.549 ± 0.051	-0.571 ± 0.052	**-0.491*** ± 0.081	-0.569 ± 0.122
Precision gap (S)	3.080 ± 0.275	3.266 ± 0.264	1.421* ± 0.181	**1.390*** ± 0.216
Independence gap (S)	0.027 ± 0.008	0.025 ± 0.004	**0.004*** ± 0.001	**0.004*** ± 0.001
Separation gap (S)	1.247 ± 0.150	1.344 ± 0.137	**0.537*** ± 0.061	0.557* ± 0.070
Sufficiency gap (S)	8.684 ± 1.883	8.816 ± 1.544	1.673* ± 0.354	**1.557*** ± 0.384

Table 7: Coreference resolution: results on Ontonotes test set and Winobias challenge set. Each model was trained over 10 seeds. * Marks significant reduction or increase in bias ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy None). The lowest bias score or highest performance metric in each column is in **bold**. P = Pearson; S = Sum.

	Debiasing Strategy			
Metric	None	Anon	CA	Anon + CA
Compression	1.984 ± 0.065	2.073* ± 0.104	**1.502*** ± 0.081	1.540* ± 0.079
F1 (Ontonotes test)	76.40* ± 0.16	76.48* ± 0.22	76.72* ± 0.15	**76.91*** ± 0.19
F1 diff ($pro - anti$)	6.072 ± 0.789	7.417* ± 1.280	3.674* ± 0.599	**2.858*** ± 0.382
TPR gap (P)	**0.635** ± 0.053	0.688* ± 0.052	0.679* ± 0.062	0.654 ± 0.049
TPR gap (S)	4.561 ± 0.414	5.143* ± 0.713	2.590* ± 0.420	**2.178*** ± 0.201
FPR gap (P)	**0.579** ± 0.046	0.637* ± 0.055	0.620* ± 0.070	0.692* ± 0.075
FPR gap (S)	0.113 ± 0.011	0.126* ± 0.016	0.063* ± 0.010	**0.052*** ± 0.004
Precision gap (P)	-0.512 ± 0.060	-0.581* ± 0.057	**-0.550*** ± 0.083	-0.632* ± 0.098
Precision gap (S)	2.943 ± 0.215	3.221* ± 0.384	1.690* ± 0.242	**1.446*** ± 0.146
Independence gap (S)	0.022 ± 0.003	0.026* ± 0.006	0.006* ± 0.002	0.004* ± 0.001
Separation gap (S)	1.188 ± 0.114	1.336* ± 0.175	0.670* ± 0.111	**0.594*** ± 0.057
Sufficiency gap (S)	7.350 ± 0.914	8.655* ± 1.726	0.2.401* ± 0.610	**1.653*** ± 0.294

Table 8: Coreference resolution after retraining: results on Ontonotes test set and extrinsic bias metrics on Winobias challenge set. Each model finetuned over 10 seeds and re-trained over 5 seeds. * Marks significant reduction or increase in bias ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy None). The lowest bias score or highest performance metric in each column is in **bold**. P = Pearson; S = Sum.

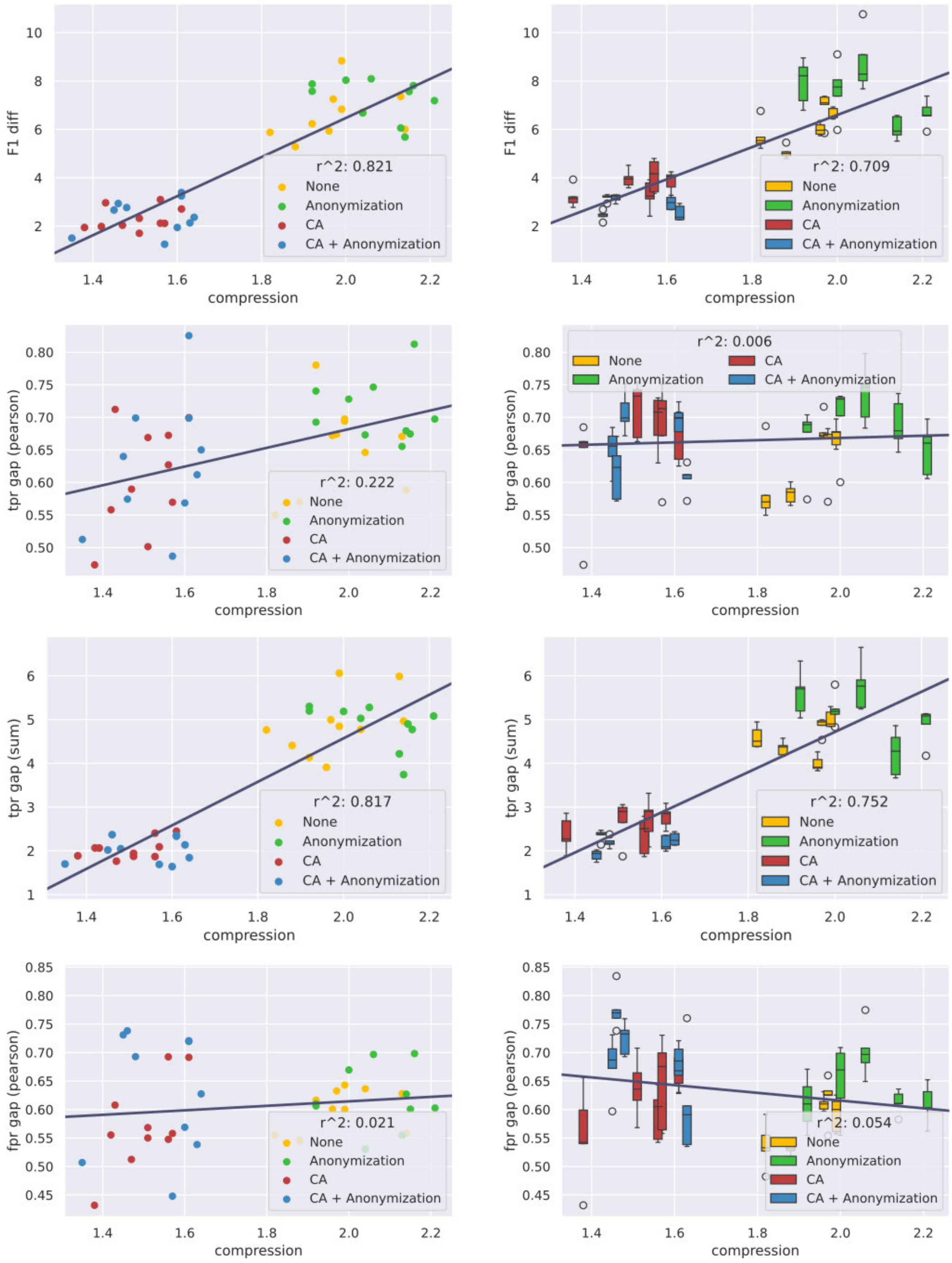


Figure 6: Coreference resolution: Before (left) and after (right) plots of compression rate versus and extrinsic metric. Cases of low and no correlation with the Pearson metrics are discussed in D.2.

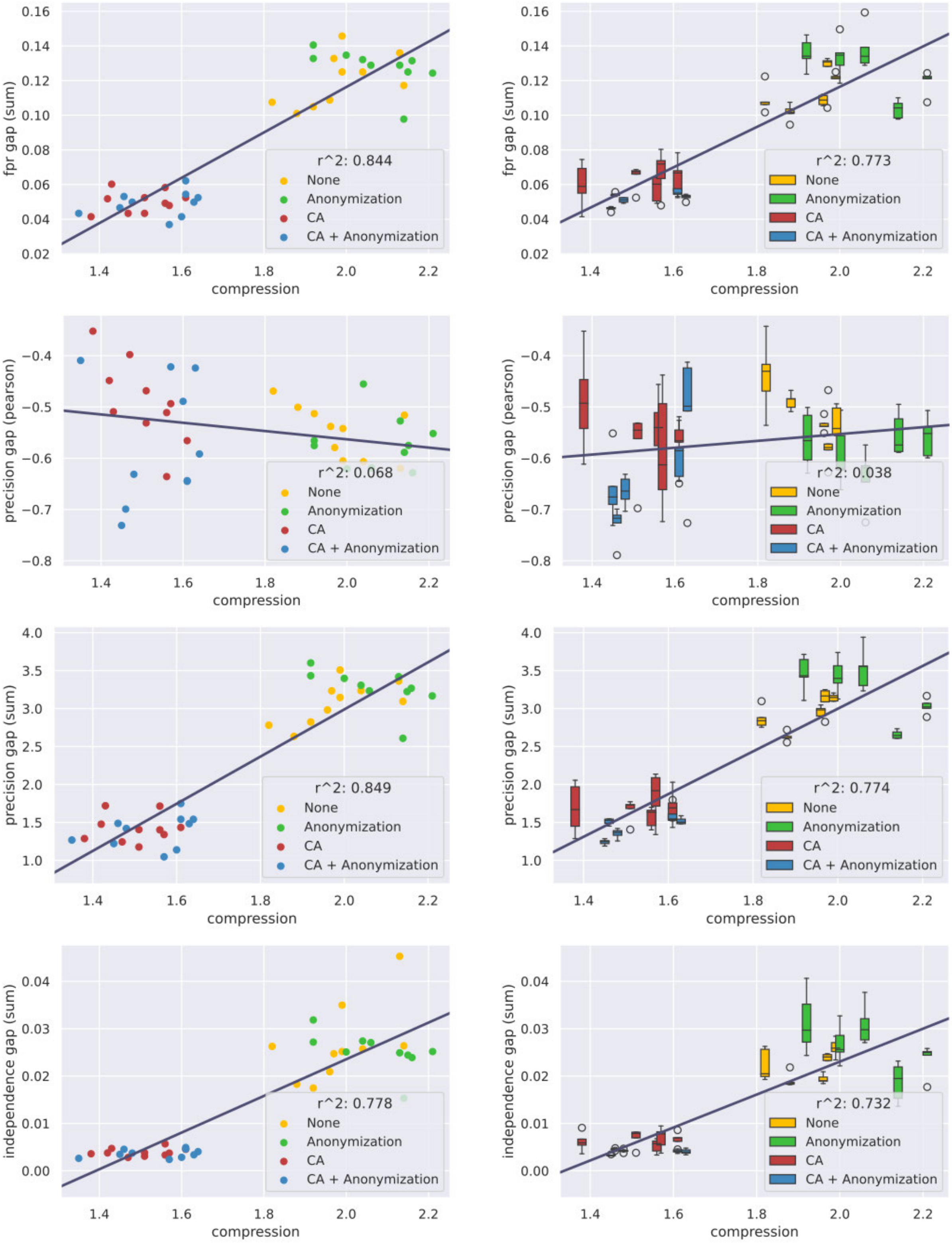


Figure 6: Coreference resolution: Before (left) and after (right) plots of compression rate versus and extrinsic metric.

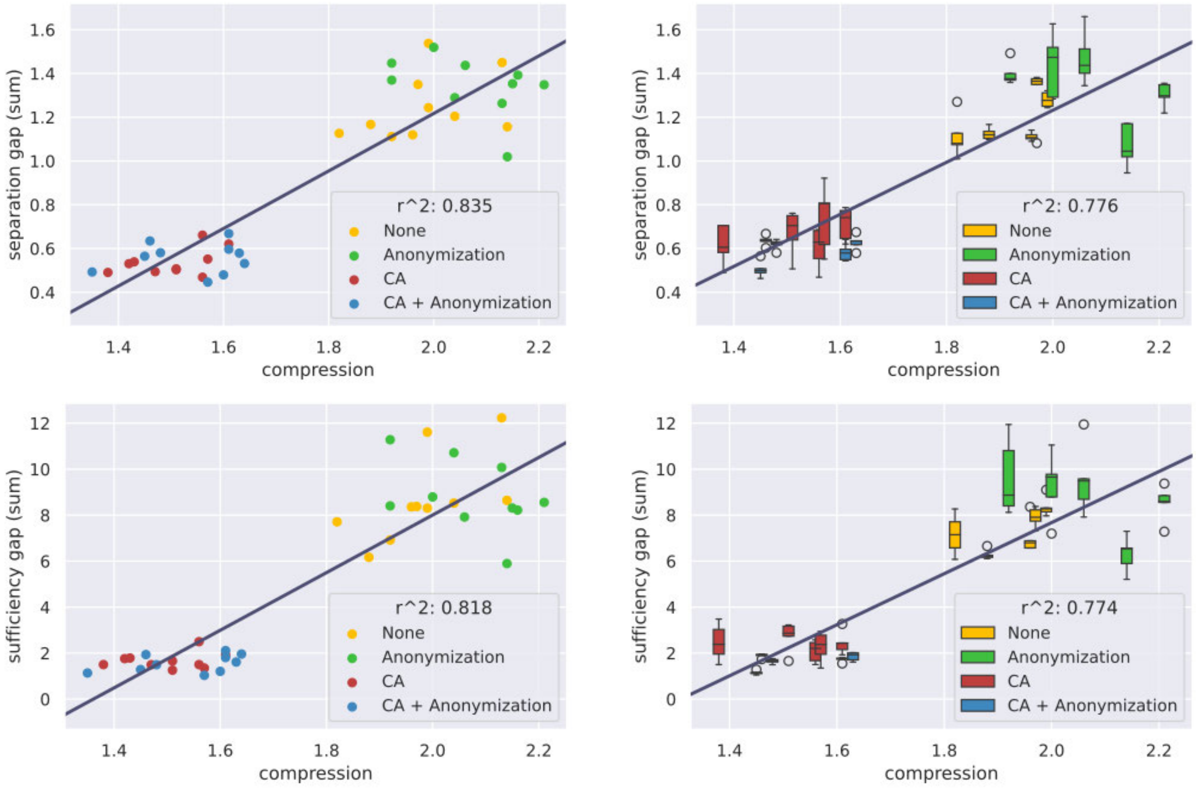


Figure 6: Coreference resolution: Before (left) and after (right) plots of compression rate versus and extrinsic metric. Cases of low and no correlation with the Pearson metrics are discussed in D.2.

Female Words	Male Words
husband, women, gender, listed, practices, nurse, specializes, children, ba, child, reading, families, location, place, affiliated, family, experiences, spanish, love, justice	chief, companies computer, applications, md, accepts, known, doctors, npi, sports, philosoph', problems, rating, no, systems, theory, practicing, software, security, major

Table 9: Top 20 significant words used to predict gender on all biographies, as obtained from a logistic regression model trained on predicting the gender of a person described in a biography. The words are sorted by importance.

Female Words	Male Words
husband , women, midwife , providing book , includes, joining, faculty	holds , emergency, vanderbilt, forces, registered, mental, assistant, president

Table 10: Top 8 words used to predict gender of female and male nurses, as obtained from a logistic regression model trained on predicting the gender of a person described in a biography. The words are sorted by importance.

C Why is scrubbing not as effective as subsampling?

The debiasing method of subsampling significantly reduced external biases in the occupation prediction task. Although compression rates show that scrubbing reduced more gender information, subsampling outperforms it as a debiasing method. We find that in spite of the scrubbing, a probe is able to correctly identify the gender from an internal representation with 68.8% accuracy compared to 90.7% on the original, non-scrubbed data. This means that although the scrubbing process reduces extrinsic bias significantly, gender information is still embedded in the [CLS] token embeddings.

To investigate the source of gender information after scrubbing, we use logistic regression (LR) model to predict the gender from the Bag-of-Words of the scrubbed biographies. We perform an iterative process for automatic extra scrubbing: in each iteration we (1) train a LR model for gender prediction (2) scrub the n most significant words for each gender according to the LR weights. The most relevant words among 5 seeds of training with n=10 words scrubbed per iteration are displayed in Table 9. The model learns indirect correlations to gender in the absence of explicit gendered words. Because the significant words are related to male- or female-dominated professions, we conducted the process on a specific profession. Table 10 presents the most significant words for biographies of nurses. There are differences in wording even between females and males in the same profession. The results of this study are in line with the results of other studies that have been conducted on the way biographies are written for men and women (Wagner et al., 2016; Sun and Peng, 2021).

Subsampling is therefore more effective even when gender information is present since it prevents the model from learning correlations between gender information and a profession whereas scrubbing only attempts to remove gender indicators without removing correlations. On the other hand, it is possible that oversampling is less effective for debiasing since seeing more non-unique examples an unrepresented group encourages learning correlations.

D A closer look into no-correlation cases

D.1 Occupation Prediction

Although compression has the ability to identify bias in most cases, some metrics still show little or no correlation with compression rate. These results suggest that gender information comprises only one facet of embedded bias in the representations. Other factors that may influence these metrics are not considered or measured, such as the connection between a name and a profession.

For example, as can be see in Tables 3 and 4, LMs finetuned on subsampled data have the largest FPR gaps after retraining, despite being the least biased before retraining, while those finetuned on oversampled data have the next-to-lowest FPR gaps after retraining. The information encoded in the internal representations may have been encoded in a manner that allowed the classification layer to exhibit a smaller FPR gap when trained on a balanced dataset. However, when the classification layer was retrained on biased training data, it used the same features to make biased predictions.

D.2 Coreference Resolution

The cases where there is no correlation between our intrinsic metric and an extrinsic metric are the cases where the metric is based on Pearson correlation. Unlike occupation prediction, coreference resolution seems to exhibit no correlation between those metrics and compression rate. These metrics are computed as the Pearson correlation between a performance gap for a specific profession and the percentage of women in that profession, however the percentages are computed differently in each task: in occupation prediction, the percentages are computed from the train set, focusing on the representation each gender has in the data. In Winobias, the percentages are taken from the US labor statistics, and are unrelated to the training dataset statistics. We note that the two statistics can be different - the real-world representation of women in a profession does not have to be equal to their representation in written text (Suresh and Guttag, 2021). We thus decided to test what happens if we change the statistics used in Winobias to dataset statistics, but Ontonotes 5.0 has very little representation to each profession and the statistics extracted from it would not be reliable. We thus took a different approach and computed the Pearson correlations for occupation prediction with real world statistics instead of dataset statistics. To do this, we mapped the professions appearing in this dataset to professions from the US labor statistics, and dropped those who could no be mapped (6 out of 29 of the professions which is 21.4%). We then repeated all experiments on the Pearson metrics using these statistics. Figure 7 shows the results. Correlations are very different when computed with respect to real-world statistics. TPR-gap has no correlation at all although it had with training data statistics, the correlation for FPR-gap after retraining exists but is negative, and the correlation with precision-gap does not exist after retraining. We thus conclude that the Pearson metrics are less reliable as they are heavily dependent on the statistics with respect to which they are calculated.

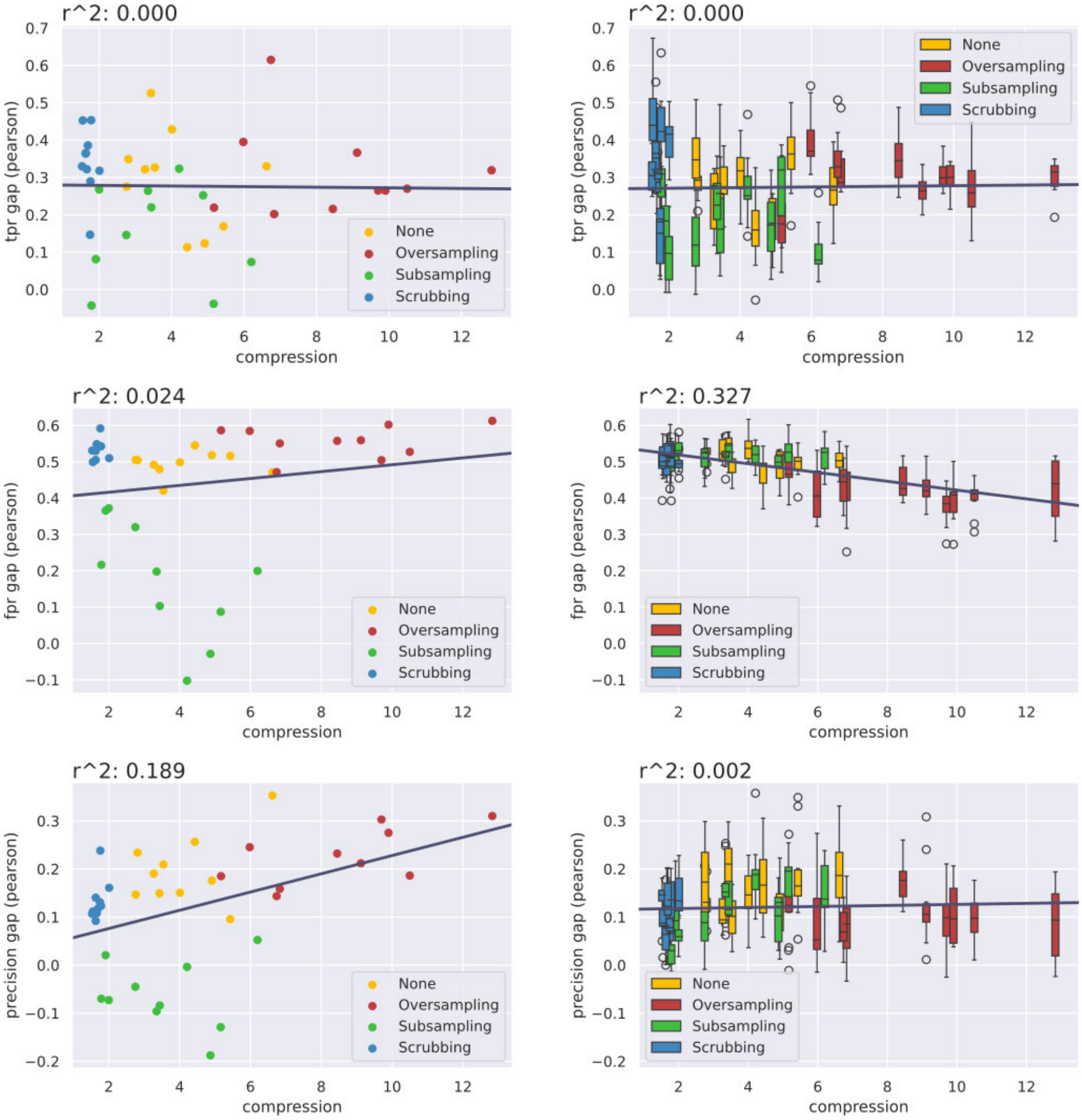


Figure 7: Occupation prediction: Before (left) and after (right) plots of compression rate versus Pearson metrics as computed from real-world statistics (as opposed to statistics derived from the training dataset). This shows the unrealiability of using real world statistics to draw conclusions, as they may not be reflected in the data.

Part II

Fairness in Transfer across Languages

We now have an understanding of the relationship between upstream and downstream fairness. In the second half of Part I we established that, within English, for gender bias, upstream potential can be realised in downstream behaviour, but is not always, depending on downstream data.

In Part II we look into more complex cases: more languages, more types of biases, and more complex transfer learning across *languages* rather than just across objectives/tasks. We ask two questions: ‘Do other languages behave like English with regard to fairness?’ ‘Does cross-lingual transfer affect fairness outcomes?’.

At the time this work was done, there were almost no resources available for testing bias downstream in multiple languages (with the exception of a multilingual version of WEAT (Lauscher and Glavaš, 2019)), which we established in Chapter 3 fails to have predictive validity). At the time of writing this thesis years later, there is still very little, despite that models are by now in the LLM-age, default multilingual. Even purportedly English only models readily speak high resource languages due to data contamination in the petabytes of pre-training data. Common use is determined by capabilities rather than by terms of use – so all models are now multilingual.

There was also a lack of pre-existing research into this topic. When this work was done there were three works examining the effect of grammatical gender on gender bias in multilingual word embeddings (Gonen et al., 2019; Zhou et al., 2019; McCurdy and Serbetci, 2017) and we built hypotheses and experiments upon this work. But there was nothing examining cross-lingual transfer. I initially proposed to look at this in 2018, so I have always found it surprising that still few have. To date there are only a handful more works.

Different languages have different distributions of strings, concepts, function words, grammatical markers, and these determine many things that matter. They determine the distribution learnt, what type of information is encoded, what is emphasised, what is compressed, what is lost. The work on multilingual word embeddings mentioned above (Gonen et al., 2019; Zhou et al., 2019; McCurdy and Serbetci, 2017) shows that grammatical gender dramatically affects representations learned even when *not* aligned to semantic gender. *La amiga* in Spanish "the female friend" has grammatical gender aligned to semantic gender, the feminine grammar form expresses the real world gender of the referent, contrasted with *el amigo* "the male friend", or in recent years also *le amige*, for expressing non-binary gender or leaving gender under-specified. *La tabla* "the table" in Spanish also expresses grammatical female gender, but is not aligned to semantic gender (as we do not generally consider tables to have semantic gender in the cultures with which I am familiar)(Fedden et al., 2018; Corbett, 1991). Those three works found that even though for most words grammatical gender has no semantic meaning, it is one of the signals most strongly encoded in representations, so it effects the learned semantics. Gonen et al. (2022) applied the PCA of Bolukbasi et al. (2016) multilingually, and found that you cannot isolate gender to the first principle component in gender marking languages, as you can in English. In gender marking languages the information is encoded across more axes.

So we know that even if we entirely discount cultural differences (which is clearly nonsense, but measuring cultural differences notoriously difficult) linguistic differences change the representations and vocabulary distributions learnt. And we also know that these same characteristics of distributions and representations: compression, what is encoded, etc, are strong causal factors in inequities and biases expressed by a model. This is what we discovered in Chapter 4. So is there some effect on fairness from superimposing one language onto another, or from blending language data? It is a more ambitious form of transfer learning with not just disjoint vocabularies (across domains) or labels (across tasks) but also *structures* (across languages). This section builds to rigorously showing that there is an effect, expressed in the second title *Cross-lingual Transfer Learning Can Worsen Biases in Sentiment Analysis*. This effect turns out to be very complex and difficult to disentangle from other confounds, but there is definitely not no effect.

In this section, we begin by creating a resource to answer our questions. We select the task of sentiment analysis as it has data in many languages. We create an evaluation benchmark to test fairness properties for sentiment analysis in a number of these languages, using the methodology for English in Kiritchenko and Mohammad (2018). Their method sets up counterfactual tests for the effect of demographics on sentiment. Once we've created the resource, we ask set of research questions that slowly build towards answering questions about cross-lingual transfer. The two works that follow directly build on each other: the first examines how transfer learning affects fairness *within one language*, for four languages (+ English). The second moves on, using the same resource, to asking more complex questions: how does cross-lingual transfer affect bias?

Models today by default use both transfer learning and cross-lingual transfer, though both things are now so common that they are generally not stated. Multilingual fairness research is not increasing commensurate with the growth of utilisation of multilingual models in practice, or even increasing much at all (Ruder et al., 2022; Blasi et al., 2022).

As with bias *analysis* as opposed to debiasing, publication processes disincentivise multilingual work. Multilingual work scales linearly in compute, experiment management, and analysis time in number of languages.

But we risk leaving other languages behind, in fairness particularly. What does it mean if an NLP system exists in one hundred languages, but is fair in only English? It is always a question whether new technologies will benefit society and improve lives, or will increase inequalities. The answer is almost always a blend of these, but quite clearly if we ensure fair NLP technologies only in English, we will tip farther towards increasing inequalities.

Chapter 5

Monolingual Transfer in Sentiment Analysis

Bias Beyond English: Count
i

Seraphina Gol
Roi Blan
†Unive

Abstract

Sentiment analysis (SA) systems are u many products and hundreds of lang Gender and racial biases are well-stud English SA systems, but understudied i languages, with few resources for such s To remedy this, we build a counterfactu uation corpus for gender and racial/n bias in four languages. We demonstrate i fulness by answering a simple but imp question that an engineer might need to a when deploying a system: What bia systems import from pre-trained models compared to a baseline with no pre-tra Our evaluation corpus, by virtue of being terfactual, not only reveals which model less bias, but also pinpoints changes in bias behaviour, which enables more ta mitigation strategies. We release ou and evaluation corpora to facilitate future research.[1]

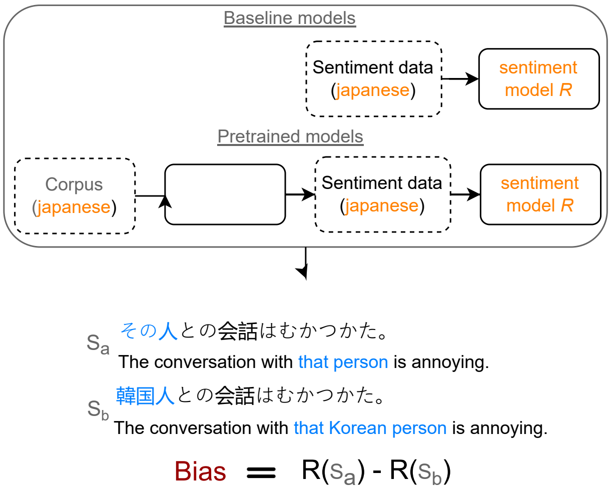

Baseline models
Sentiment data (japanese)
sentiment model R
Pretrained models
Corpus (japanese)
Sentiment data (japanese)
sentiment model R
Sa その人との会話はむかつかた。
The conversation with that person is annoying.
Sb 韓国人との会話はむかつかた。
The conversation with that Korean person is annoying.
Bias = R(Sa) - R(Sb)

Figure 1: We create corpora and then do counterfactual evaluation to evaluate how bias is transferred from training data. Counterfactual pairs (e.g. sentences *a*, *b*) vary a single demographic variable (e.g. race). We measure bias as the difference in scores for the pair. An unbiased model should be invariant to the counterfactual, with a difference of zero.

1 Introduction

Sentiment Analysis (SA) systems are among the most widely deployed NLP systems, used in hundreds of languages (Chen and Skiena, 2014). It is well-known that English SA models exhibit gender and racial biases (Kiritchenko and Mohammad, 2018; Thelwall, 2018; Sweeney and Najafian, 2020), which are acquired from their training data, training objective, and other system choices (Suresh and Guttag, 2019). Other languages are understudied; though many papers study SA bias in English, few study SA bias in other languages. This may be partly attributable to resource constraints: there are fewer corpora available to audit systems for bias in non-English languages. To remedy this, we create evaluation datasets to evaluate gender and racial bias in four languages: Japanese (ja), simplified Chinese (zh), Spanish (es), German (de). Each of these four languages has publicly available data for training SA systems (Keung et al., 2020b), and together they represent three distinct language families. To complement their existing resources with a new resource that measures bias, we use counterfactual evaluation (Figure 1), in which test examples are edited to change a single variable of interest—such as the race of the subject—extending previous work done in English (Kiritchenko and Mohammad, 2018). We release the evaluation dataset to facilitate further research.[1]

We demonstrate the value of these evaluation resources by answering the following research questions: (RQ1) What biases do we find in other languages, compared to in English? (RQ2) How does the use of pre-trained models affect bias in SA systems? While pre-trained models are common in NLP, they may import biases not present in task

[*] Correspondence to . Work completed while at an internship at Amazon.

[1] All code, evaluation data, and links to models and raw data can be found here: https://github.com/seraphinatarrant/multilingual_sentiment_analysis

supervision data, since a large pre-training corpus may embody biases not present in the supervision corpus. On the other hand, pre-training might diminish biases that arise from the small sample sizes typical of SA training corpora.

Our experiments show that both gender and racial bias are present in SA systems for all four languages: when model architecture, data quantity, and domain are held constant, SA systems in other languages display quantitatively more bias than SA systems in English. For RQ2, we find that pre-training also makes SA systems less biased for all languages, *in aggregate*, though in surprising ways: our non-pre-trained models exhibit extreme changes in behaviour on counterfactual examples, whereas pre-trained models exhibit many small nuanced changes.

2 New Counterfactual Evaluation Corpus

Counterfactual (or contrastive) evaluation establishes causal attribution by modifying a single input variable, so that any changes in output can be attributed to that intervention (Pearl, 2009). For example, if our variable of interest is gender, and our original sentence is *The conversation with that boy was irritating*, then our intervention creates the counterfactual sentence *The conversation with that girl was irritating*. Importantly, we change no other variables, such as age (*boy* → *woman*), register (*boy* → *lady*), or relationship (*boy* → *sister*). We then evaluate the behavior of our model on many such pairs of original and counterfactual sentences. In a model with no gender bias, sentiment should not change under this intervention. If it does, and does so *systematically* over many counterfactuals, we conclude that our model is biased.

To create counterfactual examples for non-English languages we use template sentences, illustrated in Table 1. Each template has a placeholder for a demographic word, in order to represent the counterfactual; and an emotion word, in order to represent different levels of sentiment polarity.

The templates of Kiritchenko and Mohammad (2018) only needed to handle the weak agreement and inflectional morphology of English, so we extend their methodology to handle a variety of grammatical phenomena in other languages. For example, in German we add gender agreement (masculine, feminine, neuter) and noun declension; in Spanish we add gender agreement (masculine, feminine, plural of both) and idiomatic verb usage;[2] in Japanese we add a distinction between active and passive forms. Chinese requires no special handling since it lacks gender agreement or inflectional morphology.

In all languages, we create a gender bias test set by providing contrasting pairs of male/female terms that can fill the placeholder for demographic variable. In German and Japanese we also provide pairs of terms for racial and anti-immigrant bias, which we derive from NGOs, sociology and anthropology resources, and government census data (Buckley, 2006; Weiner, 2009; Muigai, 2010; , FADA). We usually leave the privileged group unmarked to avoid the unnaturalness of markedness (Blodgett et al., 2021).[3] For Spanish anti-immigrant bias, we create pairs of names by using name lists that are strongly associated with migrants or with non-migrants, sourced from Goldfarb-Tarrant et al. (2021), which are based on social science research (Salamanca and Pereira, 2013). We lacked equivalent resources for Chinese, so we test only gender bias. The resulting corpora (Table 2) are comparable to or larger than other common contrastive evaluation benchmarks (Blodgett et al., 2021).

To produce the templates, we worked alongside native speakers in Japanese, German, Spanish, and Chinese to translate the English templates of Kiritchenko and Mohammad (2018), often modifying them to prefer naturalness in the target language while preserving sentiment. Our Japanese translator had professional translation experience, while our German, Spanish, and Chinese translators had training in linguistics. While collaborative development and refinement of the translation process required about a week, actual translation took about four hours for each dataset. Further details in A.

3 Methodology

For our SA task, we focus on sentiment **polarity detection** (Pang and Lee, 2007), where the output label represents the sentiment of a text as an ordinal **score** (shown in parentheses): very negative

[2] Many emotions in Spanish can idiomatically only be expressed with 'to be' or 'to have', but not both. Some take both, e.g., estoy enfadado vs. tengo un enfado — I am angry vs. I have an anger, but some emotions can use only one, or as in that example, the form changes.

[3] For example, for anti-Turkish bias in German, we replace `person dative object` in Table 1 by contrasting *dem Türken* (Turkish person (male gender)) with the unmarked *ihm* (him).

	Template	Counterfactual sentences
en	`The conversation with <person object> was <emotional situation word>.`	`The conversation with [him\her] was irritating.`
ja	`<person> との会は <emotion word passive>た`	`[彼\彼女] との会は イライラさた。`
zh	`跟 <person> 的谈话很 <emotional situation word>.`	`跟 [他\她] 的谈话很 令人生气.`
de	`Das Gespräch mit <person dat. object> war <emotional situation word>.`	`Das Gespräch mit [ihm\ihr] war irritierend.`
es	`La conversación con <person> fue <emotional situation word female>.`	`La conversación con [él\ella] fue irritante.`

Table 1: Example sentence templates for each language and their counterfactual words that, when filled in, create a contrastive pair; in this case, for gender bias. For illustration, all five examples are translations of the same sentence.

	Gender	Race/Immigrant
Japanese	3340	2004
Chinese	4928	-
German	3200	5236
Spanish	4240	6360
English	2880	5760

Table 2: Counterfactual pairs in each evaluation set, including original reference English. Differences in corpus size are due to differing number of grammatical variants and demographic words across languages.

(1), negative (2), neutral (3), positive (4), or very positive (5).[4]

3.1 Metrics

We measure the mean and variance of the differences in sentiment score between each pair of counterfactual sentences. Formally, each corpus consists of n sentences, $S = \{s_i...s_n\}$, and a demographic variable $A = \{a, b\}$ where a is the privileged class (*male* or *privileged*) and b is the minoritised class (*female* or *racial minority*). The sentiment classifier produces a score R for each sentence, and our aggregate measure of bias is:

$$\frac{1}{N}\sum_{i=0}^{n} R(s_i \mid A = a) - R(s_i \mid A = b)$$

Values greater than zero indicate bias against the minoritised group, values less than zero indicate bias against the privileged group, and zero indicates no bias. Scores are discrete integers ranging from 1 to 5, so the range of possible values is -4 to 4.

Our counterfactual evaluation process enables us to examine bias behaviour more granularly as well. We generate confusion matrices of privileged vs. minoritised scores such that an unbiased model would have all scores along the diagonal. This enables us to distinguish between many minor changes in sentiment or fewer large changes, which are otherwise obscured by aggregate metrics as described above.

In results we shade 3% of total range for easier visual inspection. This is an arbitrary choice: 'no bias' differs by application and values within the shaded range may still be unacceptable. Intuitively, this corresponds to models being maximally biased for three of every hundred examples, or making minor biased errors for twelve of every hundred.

4 Experiments

We want to answer the questions: what biases arise in SA systems in each of these languages (RQ1)? Does pre-training improve or worsen biases (RQ2)? To answer these questions, we measure the bias of a baseline SVM classification model to a model based on a pre-trained transformer model. We compare standard and distilled transformer models; distilled models are often used in practice since they are better suited to the computational constraints of real-world systems.

Our *baseline (no pre-training) models* are bag-of-words linear kernel support vector machines (SVMs) trained on the supervision data in each language. Our *pre-trained (mono-T) models* are pre-trained `bert-base` (Devlin et al., 2018) for each language. We randomly initialise a linear classification layer and simultaneously train the classifier and fine-tune the language model on the same supervision data. Our *distilled (distil-mono-T) models* are identical, but based on `distilbert-base` (Sanh et al., 2019).

We train each model five times with different random seeds (or five separate runs for the baseline) and then ensemble by taking their majority vote, a standard procedure to reduce variance. All models converge to performance on par with SotA on this task and data. Training details and F1 scores on the SA task are reported in Appendix B and C.

Training data For each model, we use the language appropriate subset of the Multilingual Amazon Reviews Corpus (MARC; Keung et al., 2020a),

[4] This is the most common approach for sentiment systems trained on user reviews, i.e. IMDB, RottenTomatoes, Yelp, Amazon products (Poria et al., 2020).

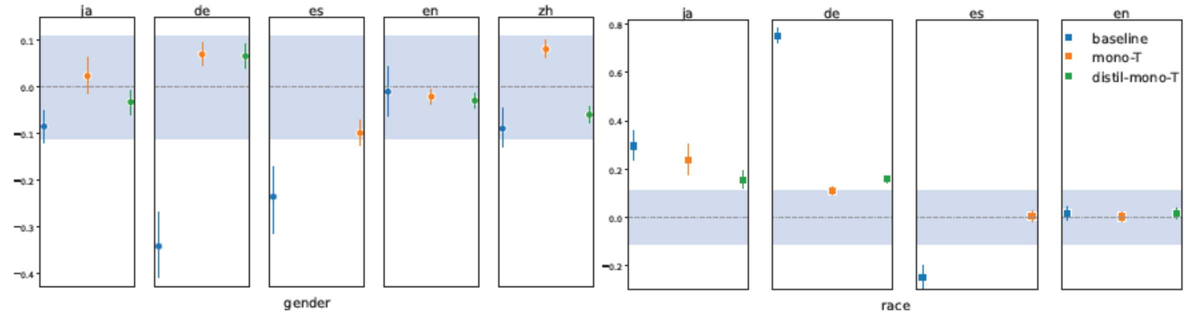


Figure 2: Aggregate bias metrics for baseline (blue), pretrained mono-T (orange), and pretrained distil mono-T (green) models. Mean and variance of differences in the sentiment label under each counterfactual pair, one graph per language and type of bias tested. Higher numbers indicate greater bias against the minoritized group. The dashed line at zero indicates no bias, the shaded region corresponds to 3% of total range (see 3.1). Spanish (es) distilled model is intentionally missing for lack of comparable pretrained model.

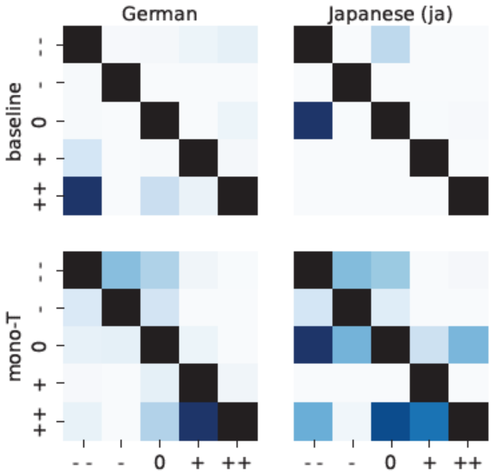


Figure 3: Confusion matrices for racial counterfactual pairs for Japanese and German, comparing baseline and pretrained models. Higher colour saturation in the lower triangle is bias against the minoritised group, against the privileged group in the upper triangle.

which contains 200 word reviews in English, Japanese, German, French, Chinese and Spanish, with discrete sentiment labels ranging from 1-5, balanced across labels.

5 Results

The baseline models are most biased for both gender and race in all languages (Figure 2), though not always *against* minoritised groups: systems are often biased against the male demographic, consistent with previous work on SA (Thelwall, 2018). [5]

Figure 2 also shows that English models tend to be less biased than the other languages.

Analyzing the granular differences (Figure 3) reveals interesting behaviour not captured by aggregate metrics: much of the bias exhibited by the baselines arises from consistently flipping *specific* labels in the counterfactual, while bias exhibited by pre-trained models is more varied.[6] For example, the Japanese baseline exhibits racial bias by frequently changing neutral labels to very negative labels, whereas in the mono-T model the change under the counterfactual is expressed as many less extreme changes. The model is still biased overall: though the changes are more varied, in aggregate they associate racial minorities with more negative sentiment. The German baseline model is more extreme: when the demographic variable changes from privileged to minoritised, the model changes its prediction from very positive to very negative. The German mono-T model also makes biased choices, though more moderately (neutral to negative) and there is more 'counter-bias' in the upper triangle, which lessens overall bias.

6 Related Work and Conclusion

Counterfactual evaluation is frequently used in bias research on classification tasks (Garg et al., 2019), and sometimes even on generation tasks (Huang et al., 2020). There have also been works exposing common pitfalls in the design of counterfactuals (Blodgett et al., 2021; Zhang et al., 2021; Krishna et al., 2022). Anyone expanding or replicating our counterfactual evaluation work should consult these as prerequisites. The contemporary work of

[5] Because this task is sentiment analysis, it is more possible to get bias against a male demographic than if the task were, say, biography classification. For the latter, the male demographic is associated with prestige roles (and thus generally bias is anti-female), but for sentiment analysis, male demographics can be associated with negative characteristics (violence, aggression, if a model is stereotyping) as well as with competence, so a few works have found female subjects to sometimes have more positive sentiment, depending on context.

[6] We show Japanese and German for illustration; the trend is present in all languages. All graphs are in Appendix D.

Seshadri et al. (2022) find many ways that other templates for bias evaluation can be brittle, so future work should take this into account and take measures to ensure robustness, such as testing with multiple paraphrases of the templates.

We have laid the groundwork for investigating bias in sentiment analysis beyond English. We created resources, presented an evaluation procedure, and used it to do the first analysis of bias in SA in a simulated low-resource setting across multiple languages. We showed that using pre-trained models produces *much less* biased models than using baseline SVMs. We also showed that pre-trained models have very different *patterns* of bias; a type of analysis that is enabled by the counterfactual design of our corpus. We invite the NLP community to use the data and methods from this work to continue analysis of languages beyond English.

7 Limitations

Like all bias tests, these experiments have *positive* predictive power: they can find the biases they test for, but they cannot eliminate the possibility of there being biases that the tests overlook.

Our Japanese, German, Spanish, and Chinese translators were from Japan, Germany, Spain, and mainland China, respectively. Hence, their translations may reflect their native dialects of these languages. While these dialects are consistent with the corresponding training datasets in these languages, this fact may limit conclusions that we or others can draw about SA in other dialects of these languages, such as Central and South American dialects of Spanish, or Chinese (Traditional).

8 Ethics Statement

Because of the aforementioned limitation regarding positive predictive power, there is always a risk with research on social biases that it can give practitioners a false sense of security. It is absolutely possible to evaluate on our corpus and get no bias, and still end up causing harm to racial or gender demographics, since they do not cover all biases or all domains. This should be kept in mind whenever applying this research.

Acknowledgements

We thank Björn Ross for many comments and helping shape the draft, Lluís Màrquez for helping manage the project at Amazon, and the Amazon Barcelona Search team for their enthusiastic support of the project.

References

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Sandra Buckley. 2006. *Encyclopedia of contemporary Japanese culture*. Routledge.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

The Federal Anti-Discrimination Agency (FADA). 2020. Equal rights, equal opportunities: Annual report of the federal anti-discrimination agency.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 219–226, New York, NY, USA. Association for Computing Machinery.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020a. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Phillip Keung, Julian Salazar, Yichao Lu, and Noah A. Smith. 2020b. Unsupervised bitext mining and translation via self-trained contextual embeddings. *Transactions of the Association for Computational Linguistics*, 8:828–841.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Satyapriya Krishna, Rahul Gupta, Apurv Verma, Jwala Dhamala, Yada Pruksachatkun, and Kai-Wei Chang. 2022. Measuring fairness of text classifiers via prediction sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5830–5842, Dublin, Ireland. Association for Computational Linguistics.

Githu Muigai. 2010. Report of the special rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, githu muigai, on his mission to germany (22 june - 1 july 2009).

Bo Pang and Lillian Lee. 2007. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *CoRR*, abs/2005.00357.

Gastã Salamanca and Lidia Pereira. 2013. PRESTIGIO Y ESTIGMATIZACIÃ“N DE 60 NOMBRES PROPIOS EN 40 SUJETOS DE NIVEL EDUCACIONAL SUPERIOR. *Universum (Talca)*, 28:35 – 57.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

Harini Suresh and John V. Guttag. 2019. A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002.

Chris Sweeney and Maryam Najafian. 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 359–368, New York, NY, USA. Association for Computing Machinery.

Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review*.

Michael Weiner. 2009. *Japan's minorities: the illusion of homogeneity*, volume 38. Taylor & Francis.

Chong Zhang, Jieyu Zhao, Huan Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021. Double perturbation: On the robustness of robustness and counterfactual bias evaluation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3899–3916, Online. Association for Computational Linguistics.

A Benchmark Dataset Creation

We followed the recommendations of Blodgett et al. (2021) to ensure the validity of our datasets. Many of the pitfalls enumerated in their work do not apply to our dataset, as we are measuring sentiment, rather than stereotypes, but we took care to avoid those that do apply. These are:

Markedness. In most cases we contrast the minority group, e.g. *Turkish people* with the unmarked group, e.g. *people*. Using a marked privileged group—white people, straight people, etc—is in most cases uncommon and occurs in only particular settings, which threatens the validity of the contrastive test (Blodgett et al., 2021). We do make a few exceptions and mark privileged groups. We do mark them for gender bias, since gender is explicitly marked in language more than other demographic traits (e.g. we contrast *woman* with *man*, not with *person*). We also sometimes use first names as proxies for demographics such as race, class, and immigration status (in Spanish and English) and in these cases the privileged group is another name.

Naturalistic Text. Some of the sentences in the original Kiritchenko and Mohammad (2018) would be valid grammatical sentences if translated directly into other languages, but would not sound natural. For example, reflexive pronouns (himself, herself) aren't used the same way in Chinese as in English, so in translating the English template <person subject> found himself/herself in a/an <emotional situation word> situation. we instead used the Chinese template <person subject> 经历了一件<emotional situation word> 的事., which means <person subject> was in a <emotional situation word> situation. These small changes preserve the same rough semantics, and more importantly preserve naturalness.

Indirect Demographic Identification. Blodgett et al. (2021) caution against the use of proper names or other proxies as a stand in for a demographic group, because their reliability for this use is untested. We would add that names are difficult to use in a contrastive pair where we need to change only *one* demographic variable, because names indicate many bits of demographic information at once: race, gender, class, place of birth, period of birth, etc. We intentionally avoid this by using identity terms (Turk, Korean, etc) most of the time, which do sometimes conflate race and country of origin, but are otherwise the most precise option. We use proper names only in Spanish based on the work of Goldfarb-Tarrant et al. (2021) and Salamanca and Pereira (2013), who show that there is data backing up the migrant vs. non-migrant names. Even so, there is some conflation between migrant status and socioeconomic class in that set of names: we consider that acceptable for our purposes. There are also names as a proxy for African-Americans in English, as the dataset is from Kiritchenko and Mohammad (2018) and that is what they use.

Basic Consistency A few other applicable pitfalls, which Blodgett et al. (2021) capture under the heading 'Basic Control and Consistency' we avoid organically by our template based construction, e.g. differences in sentence length between sentences A and B, are a possible confound, but by construction we contrast only one word in a pair and the sentence is otherwise unperturbed.

Once we had designed our translation process, we did a multi-step qualitative evaluation. After we had settled on the first version of the three sets of templates, demographic terms, and emotion words in each language, we worked with the native speaker to iterate and make sure there were no accidental unnatural sentences or grammatical errors. We generated a few examples for each template + emotion + demographic combination, manually reviewed 200 examples per language, and then made corrections to the templates, words and the rules for combining them. We then repeated this exact process a second time after the adjustments.

B Model Implementation Details

Monolingual transformer models have 110 million parameters (± 1 million) and vocabularies of 30-32k with 768D embeddings. We train the monolingual models with the same training settings as preferred in Keung et al. (2020a), and allow the pretrained weights to fine-tune along with the newly initialised classification layer.

C Model Performance

Performance at convergence for models in each language is given in Table 3.

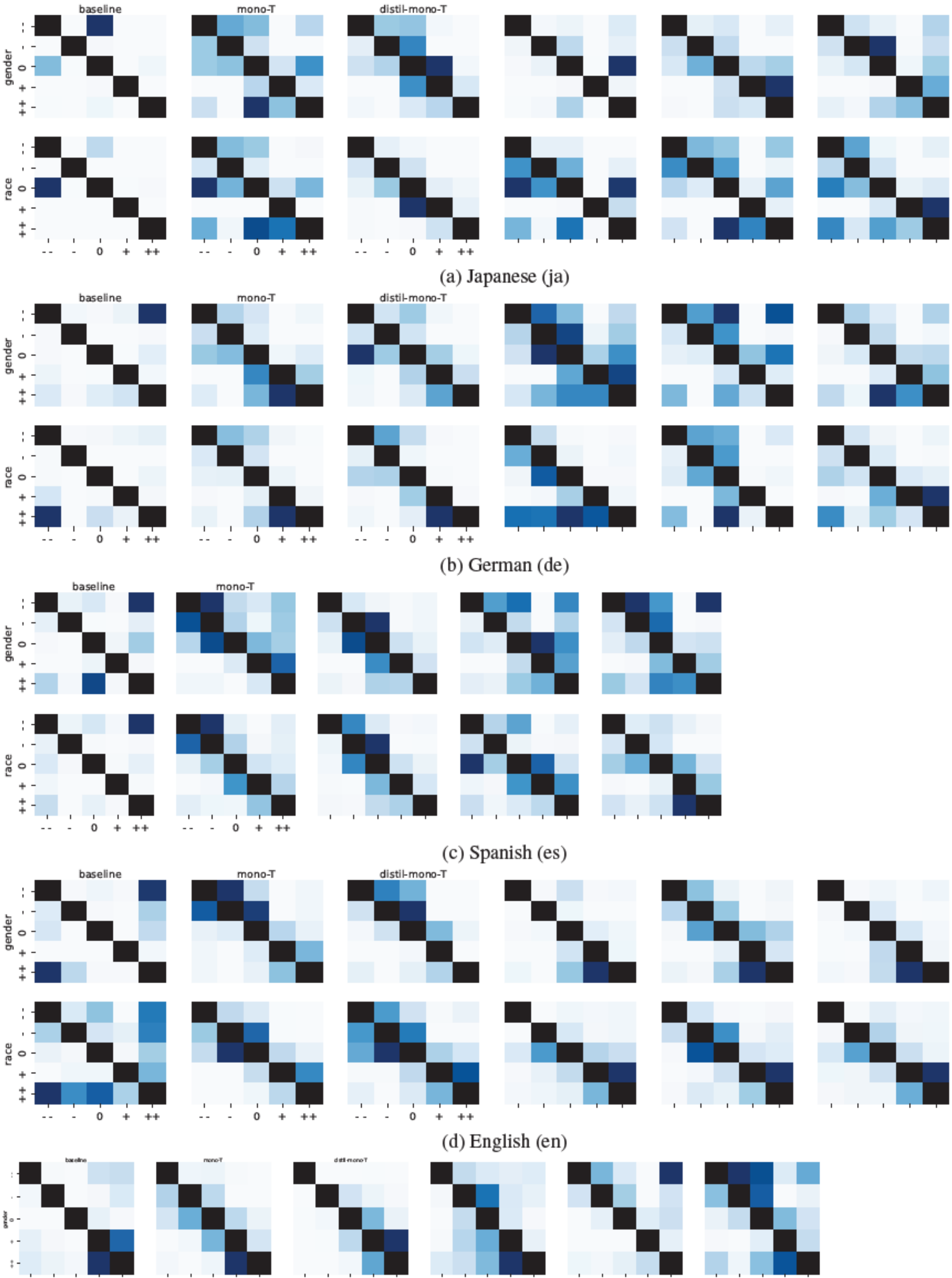
We determined convergence by examining loss curves and selecting the model where training loss was flat, and validation had not yet increased. We did not use early-stopping, as we wanted to save many model checkpoints in order to study the training dynamics of bias, including *after* convergence when the model was overtrained. However, we found no clear trends in how bias changed over the course of training, so for this study we used only one model, at convergence, per language. We hope that by releasing all model checkpoints (15 per language), other researchers may be able to expand our work into the training dynamics of bias.

	Standard		**Distilled**		**Baseline**
	F1	Steps	F1	Steps	F1
ja	**0.62**	44370	0.61	60436	0.38
zh	**0.56**	35190	0.53	43750	0.42
de	**0.63**	36720	0.63	52621	0.51
es	**0.61**	41310		-	0.48
en	**0.65**	27050	**0.65**	44285	0.53

Table 3: F1 at convergence and steps at convergence for standard size, distilled, and baseline models. Performance is measured on the MARC data.

D Full set of confusion matrices comparing baseline and monolingual models.

Figure 4 contains all confusion matrices for all languages, of which we displayed a subset in the body of this work.



(a) Japanese (ja)

(b) German (de)

(c) Spanish (es)

(d) English (en)

(e) Chinese (zh_cn)

Figure 4: All confusion matrices for experiments in this paper. Higher colour saturation in the lower triangle is bias against the minoritised group, in the upper triangle is bias against the privileged group. Saturations are not normalised across all languages and models; this is not a proxy for aggregate comparative bias, it shows the pattern across sentiment scores.

Chapter 6

Cross-lingual Transfer in Sentiment Analysis

The next work directly follows on from the previous results: the experiments were planned together and directly inform each other, despite being published separately.

In the previous work, we created the resources needed to do these experiments, as it was the first work on fairness in language models across multiple language families. We found that there is an effect on fairness from transfer learning within one language. We can't disentangle the exact *causes* of this effect from those experiments: whether it is information contained in the data, or the additional stability of the model from the addition of more data, though we hypothesise the latter (stability from more data) is the cause. Regardless of the causes, the findings are useful in practice under resource constraints, if less scientifically satisfying than if we had controlled all variables.

In the following work, we examine the more complex setting of cross-lingual transfer in all the same languages, and again ask how this setting changes fairness outcomes. However, we set up our experiments to control as many variables as possible and establish causes, without pre-training all new models from scratch (that is, we limit ourselves to fine-tuning only, as the multilingual setup has already extremely many variables and requirements on compute resources). The full set of experimental variables we consider is:

- Type of bias. We look at gender bias and racial/country of origin bias. We might expect these to have different patterns of cross-lingual transfer as gender is encoded in some languages in a way that race is not (via gender agreement) and as gender biases tend to be global and common across languages in a way that racial biases are not (the minoritised racial groups differ culture to culture).

- Mono vs. multilingual pre-training. We examine what happens when changing from monolingual to multilingual pretraining *without* changing the fine-tuning data. This would not be none in practice in a production system, but enables us to isolate the two types of data (pretraining and fine-tuning) that usually change when going from a monolingual transfer to a cross-lingual transfer model. In our first experiments, we hold fine-tuning data constant for each language and change only pretraining data.
- Target language fine-tuning vs. transfer language fine-tuning. In a monolingual transfer setup, a model applied to Spanish as the target language will be fine-tuned on Spanish. In cross-lingual transfer, it will be fine-tuned in another language (in this case English) and then applied to Spanish. In these next experiments, we hold the pretrained multilingual model constant and change the fine-tuning data.
- Random seed. All experiments report the majority vote over five random seeds for the weight initialisation of the classifier and the data shuffle for fine-tuning. We initially did an analysis by individual random seed, and found them to differ so strongly that sometimes even polarity of the bias flipped: that is, for random seed A there would be anti-female bias and for random seed B there would be anti-male bias. We take majority vote to indicate what would be the most likely thing to happen for a random seed picked out of a hat.
- Distillation. We do the same set of experiments for full-size (100-150 million parameters) and for distilled models, which are approcimately half the number of parameters. These experiments make our results more applicable in practice, as distilled models are commonly used in combination with cross-lingual transfer as both are methods to deal with insufficient data or resources.

There are nonetheless two experimental variables that we consider important but were unable to include. We do not look at the effect of modifying pretraining data: it is an important variable in the manifestation of social bias, but it is the most difficult to experiment on because training a model from scratch is so challenging. It is also the one least likely for developers of NLP systems to modify in practice, for that very reason. We also do not look at the effect of domain match/mismatch. In practice many sentiment systems have a domain mismatch, since sentiment training data tends to be from domains where sentiment can be determined from freely available metadata without an annotation effort: movie, restaurant, and product reviews. Our experiments reflect this domain mismatch by training on product reviews and using standard text at inference. However, results may differ for in-domain data, and it would be possible to also create a bias evaluation dataset from in-domain data and observe the differences.

As a result of these ablations, this work focuses on two types of causal tests. The evaluation dataset is based on counterfactual pairs, which are causal tests that answer the question not just *what changed* (as an observational study does) but also *why did it change* (a property of an interventional or causal study). This is a different *why* than is answered when we do all of our experiments, which are themselves a different kind of counterfactual. The evaluation data holds everything fixed save the demographic variable, such that any change is attributable to the perturbation of the demographic variable. The many experimental scenarios hold everything fixed save a specific difference in model training, so that difference becomes the variable that can establish causality. We also leverage the analytical method from the previous work on using a counterfactual confusion matrix to visually inspect patterns of bias.

The many experiments thus answer whether the observed behaviour came from pretraining or fine tuning as best as possible. There is a limitation, which is that this setup is unable to isolate any interaction effects (which there almost certainly is because pretraining sets inductive biases). It also doesn't answer what about each step caused the change (what segment of data, what hyperparameter). We are unaware of any work that can manage these questions, but we do want to call out that though this work is very rigorous on causal attribution, it is still able to establish causality to only a limited extent and far more research into this area is needed for it to be understood. We released all the models in the hopes that other researchers will do some of this work.

Cross-lingual Transfer Can Worsen Bias in Sentiment Analysis

Seraphina G
School of
University

C

Sentiment ana
ployed in man
there is well-
graphic bias i
beyond Englis
supplemented
trained model
els trained on
even supervis
guages. Does
new biases? T
counterfactual
der or racial b
cross-lingual t
gual transfer s
find that syste
usually becom
gual counterpa
be much more
spur further re
the sentiment
and the intern
training, yiel
also release ou

1 Introductio

Sentiment analys
cations, leading
for many languag
pervised learning
of supervised trai
languages. Since
in a new languag
strategies are cor
even to avoid it a
cost, is **monoling**
supervised mode
language, fine-tu
sion data in that
in that language

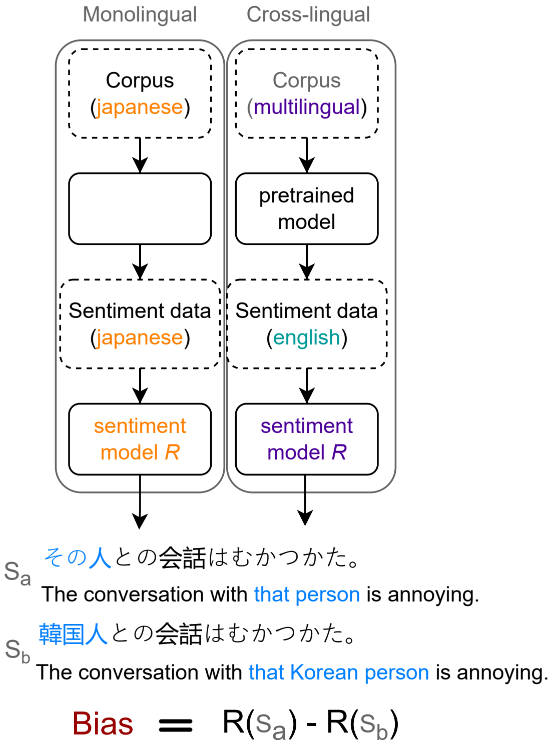

Monolingual
Cross-lingual
Corpus (japanese)
Corpus (multilingual)
pretrained model
Sentiment data (japanese)
Sentiment data (english)
sentiment model R
sentiment model R
Sa
その人との会話はむかつかた。
The conversation with that person is annoying.
Sb
韓国人との会話はむかつかた。
The conversation with that Korean person is annoying.
Bias = R(Sa) - R(Sb)

Figure 1: We use counterfactual evaluation to evaluate how bias is differs in monolingual vs. cross-lingual systems. Counterfactual pairs (e.g. sentences *a*, *b*) vary a single demographic variable (e.g. race). We measure bias as the difference in scores for the pair. An unbiased model should be invariant to the counterfactual, with a difference of zero.

second, which avoids annotation cost altogether, is **zero-shot cross-lingual transfer**: we pre-train an unsupervised model on a large corpus in *many* languages, fine-tune on already available supervision data in a high-resource language, and use the model directly in the target language (Eisenschlos et al., 2019; Ranasinghe and Zampieri, 2020).

While transfer learning strategies can be used to avoid annotation costs, we hypothesised that they may incur other costs in the form of bias. It is well-known that high-resource SA models exhibit gender and racial biases (Kiritchenko and Mohammad, 2018; Thelwall, 2018; Sweeney and Najafian,

[1]https://git
multilingual_s

2020). Less is known about bias in other languages. A recent study found that SA models trained with monolingual transfer were less biased than those trained without any transfer learning (Goldfarb-Tarrant et al., 2023). As far as we are aware, there is no work that studies the effect of cross-lingual transfer on bias.

But there is good reason to hypothesise that cross-lingual transfer may introduce new biases. Specific cultural meanings, multiple word senses, and dialect differences often contribute to errors in multilingual SA systems (Mohammad et al., 2016; Troiano et al., 2020), and are also sources of bias (Sap et al., 2019). For example, the English word *foreigner* translates to the Japanese word *gaijin* (外人) which has approximately the same meaning, but more negative sentiment. Bias may also arise from differences in what is explicitly expressed. For example, there is evidence that syntactic gender agreement increases gender information in representations (Gonen et al., 2019a; McCurdy and Serbetci, 2017), and there is also evidence that gender information in representations correlates with gender bias (Orgad et al., 2022). From these facts, we hypothesise that multilingual pre-training on languages with gender agreement will produce more gender bias in target languages without gender agreement, while producing less bias in target languages with gender agreement.

In this paper, we conduct the first investigation of biases imported by cross-lingual transfer, answering the following research questions: **(RQ1)** What biases are imported via cross-lingual transfer, compared to those found in monolingual transfer? When using cross-lingual transfer, are observed biases explained by the pre-training data, or by the cross-lingual supervision data? Since practical systems often use distilled models, we also ask: **(RQ2)** Do distilled transfer models show the same trends as standard ones?

We investigate these questions via counterfactual evaluation, in which test examples are edited to change a single variable of interest—such as the race of the subject—so that any change in model behaviour can be attributed to that edit. We use the counterfactual evaluation benchmarks of Kiritchenko and Mohammad (2018) and an extension of it (Goldfarb-Tarrant et al., 2023) to test for gender, racial, and immigrant bias in five languages: Japanese (ja), simplified Chinese (zh), Spanish (es), German (de), and English (en). The first four languages cover three different language families, that all have fewer sentiment analysis resources then English; including English in the study enables us to compare to previous work. We find that:

1. Zero-shot multilingual transfer generally increases bias compared to monolingual models. Racial bias in particular changes dramatically.

2. The increase in bias in cross-lingual transfer is largely, but not entirely attributable to the multilingual pre-training data, rather than cross-lingual supervision data.

3. As hypothesised, gender bias is influenced by multilingual pre-training in directions that are predictable by the presence or absence of syntactic gender agreement in the target language.

4. Compressing models via distillation often reduces bias, but not always.

We conclude with a set of recommendations to test for bias in zero-shot cross-lingual transfer learning, to create more resources to allow testing, and to expand bias research outside of English. We release all models and code used for our experiments, to facilitate further research.[1]

2 Background

2.1 Cross-lingual Transfer

The aim of transfer learning is to leverage a plentiful resource to bootstrap learning for a task with few resources. Cross-lingual transfer learning (Ruder et al., 2019; Pires et al., 2019; Wu and Dredze, 2019) extends this idea to transferring across *languages*. It works by pre-training a model on text in many languages, including both the target language and one or more additional languages with substantial resources in the target task. For example, we pre-train a model on a multilingual web crawl containing both English and Japanese, and fine-tune on many English reviews (plentiful resource). We then assume that since the model knows about *both* Japanese and polarity detection, it can be applied to the task even though it has never seen examples of polarity detection in Japanese. We call this zero-shot cross-lingual transfer (**ZS-XLT**). An alternative approach is few-shot transfer, where we also use a very small amount of target-language supervision. We focus on zero-shot transfer because it makes clear any causal link between multilingual training and bias transfer.

2.2 Counterfactual Evaluation

Counterfactual evaluation is an approach that allows us to establish causal attribution: a single input variable is modified at a time, so that one can be sure that any changes in the output are due to that change (Pearl, 2009).

Benchmarks for evaluating model fairness with this strategy are constructed so that model predictions should be invariant to changes in a demographic or protected variable such as race or gender (Kusner et al., 2017).[2] For example, the sentiment scores of *The conversation with that boy was irritating* and *The conversation with that girl was irritating* should be equal. If there is a systematic difference in predicted sentiment scores between such pairs of sentences, we conclude that our model is biased. Biased models for sentiment analysis are likely to propagate representational harm (Crawford, 2017) by systematically associating minoritised groups with more negative sentiment. They also can propagate allocational harm by being less stable at sentiment prediction in the presence of certain demographic information. Sentiment analysis is often a component of another application, so the specific harm depends on the application.

3 Methodology

We treat sentiment polarity detection as a five-way classification problem: very negative (1), negative (2), neutral (3), positive (4), or very positive (5). In figures, we refer to these classes by using symbols **--**, **-**, **0**, **+** and **++**. This ordinal labeling scheme is commonly used when systems are trained on user reviews with a star rating (Poria et al., 2020).

We train monolingual and cross-lingual models, then evaluate them on counterfactual corpora and compare their differences in bias measures. We look at both average bias using aggregate metrics and granular bias using a contingency table of counterfactuals. This enables us to build an overall picture of model comparability and also to differentiate between models with identical aggregate bias but different behaviour – some models may make many small errors, and some may make few large errors, and this may matter for minimising real world harms.

3.1 Evaluation Benchmarks

To evaluate social bias in our experiments, we use multiple different counterfactual benchmarks. Table 1 contains examples from all datasets. For English, we use the counterfactual corpus of Kiritchenko and Mohammad (2018), which covers binary gender bias, and racial bias. Gender is represented by common gender terms (*he*, *she*, *sister*, *brother*), and African American race is represented by African-American first names contrasted with European American ones, derived from Caliskan et al. (2017). For non-English language benchmarks, we use the corpus of Goldfarb-Tarrant et al. (2023) which follows the methodology of Kiritchenko and Mohammad (2018) to create the same kind of benchmark in German, Spanish, Japanese, and Chinese, extended to respect linguistic and cultural specifics of those languages. In the Goldfarb-Tarrant et al. (2023) benchmark, all languages have a test for gender bias, where gender is binary and is similarly represented by common gender terms (as above). The German resource covers anti-immigrant bias, using identity terms of race and nationality identified by governmental and NGO resources as immigrant categories that are targets of hate (Muigai, 2010; , FADA) e.g. *Turk, Arab, Muslim, Roma, Sinti*. The Japanese resource covers bias against racial minorities, using identity terms of minoritised groups from sociology resources (Buckley, 2006; Weiner, 2009), e.g *Chinese, Korean, Okinawan*. The Spanish resource tests anti-immigrant bias via name proxies of immigrant first names, taken from Goldfarb-Tarrant et al. (2021) based on the social science research of Salamanca and Pereira (2013). The benchmark provides only gender bias tests for Chinese, so this work includes an analysis of gender bias only for Chinese. For reference, we have included the full set of racial and nationality groups covered in the benchmark in Appendix C.

In all datasets, counterfactual pairs are generated from template sentences (Table 1) that vary both the counterfactual and the sentiment polarity, by using placeholders for demographic words and emotion words, respectively. Demographic words are as described above, for emotion words, Kiritchenko and Mohammad (2018) use 40 English emotion words that fit into high level categories of fear, anger, joy, and sadness (this granularity allows testing more granular sentiment and emotion rather than simply polarity, if desired). Goldfarb-Tarrant et al. (2023)

[2] There are tasks where invariance to demographics doesn't make sense, such as hate speech classification. Our evaluation data are designed so that all examples should be invariant.

use emotion words from the same high level categories and about 10 emotions per category as well, though sometimes this is many more than 10 actual words to account for grammar in non-English languages (gender, case, etc). Datasets range from 3-5k pairs per language, which gives sufficient statistical power for the differences we observe. We nonetheless include confidence intervals in all our analysis.

There exists an additional benchmark of the same construction covering Arabic also Câmara et al. (2022). It was not yet available at the beginning of this work (and there is no equivalent Arabic sentiment data for us to use) so we did not use it in this work, but it may be helpful for future research to include an additional language family.

3.2 Metrics

We need an aggregate measure of overall bias and a way to look at results in more detail. For our aggregate metric, we measure the difference in sentiment score between each pair of counterfactual sentences, and then analyse the mean and variance over all pairs. Formally, each corpus consists of n sentences, $S = \{s_i ... s_n\}$, and a demographic variable $A = \{a, b\}$ where a is the privileged class (*male* or *privileged / unmarked race*) and b is the minoritised class (*female* or *racial minority*). The sentiment classifier produces a score R for each sentence, and our aggregate measure of bias is:

$$\frac{1}{N}\sum_{i=0}^{n} R(s_i \mid A = a) - R(s_i \mid A = b)$$

In this formulation, values greater than zero indicate bias against the minoritised group, values less than zero indicate bias against the privileged group, and zero indicates no bias. Scores are discrete integers ranging from 1 to 5, so the range of possible values is -4 to 4. For example, if a sentence received a score of 4 with the male demographic term, and a score of 1 with the female demographic term, then the score gap for that example is 3.

To put our results in context, Kiritchenko and Mohammad (2018) found the average bias of a system to be $\leq 3\%$ of the output score range, which corresponds to a gap of 0.12 on our scale. In practice, this is equivalent to reducing the sentiment score by one for twelve out of every hundred reviews mentioning a minoritised group, or to flipping the score from maximally positive to maximally negative for three out of every hundred.

For more granular analysis we examine contingency tables of privileged vs. minoritised scores for each example. This enables us to distinguish between many minor changes in sentiment or fewer large changes, which are otherwise obscured by aggregate metrics as described above.[3]

4 Experimental Setup

Our goal is to simulate practical conditions as much as is possible with available resources and datasets, so we start with pre-trained models from huggingface (Wolf et al., 2020) which are commonly used in sentiment benchmarks and previous work on our data.[4] We then fine-tune these models on supervised training for the polarity detection task and apply to the counterfactual evaluation set in the target language. Both monolingual and multilingual models have as similar numbers of parameters and fine-tuning procedures as is possible, to minimise confounds while being realistic (Appendix A). Models are fine-tuned until convergence using early stopping on the development set. All models (multilingual and monolingual) converge to equivalent performance as previous work (Keung et al., 2020), which is state of the art on this task. F1 scores and steps to convergence are included in Appendix B.

Monolingual transfer (mono-T) models are based on pre-trained `bert-base-uncased` (Devlin et al., 2018) in the target language. We randomly initialise a linear classification layer, then simultaneously train it and fine-tune the language model on monolingual supervision data. Our distilled monolingual model (**distil-mono-T**) is identical, except that it is based on `distilbert-base-uncased` (Sanh et al., 2019).

Multilingual models are based on pre-trained `mbert-base-uncased` then fine-tuned on a large volume of sentiment data in English only, the standard approach to zero-shot cross-lingual transfer **(ZS-XLT)**. We also fine-tune a distilled ZS-XLT model (**distil-ZS-XLT**), identical except that it is based on `distilmbert-base-uncased`.

[3] Readers familiar with Kiritchenko and Mohammad (2018) may recall that they provide an aggregate measure in the form of a graph, as we do, and more granular measures of amount of bias per group (e.g. for male and female separately), in a table. We forgo the table as we use contingency tables in our analysis, which contain a superset of the same information (bias by group, as well as bias by label).

[4] https://paperswithcode.com/task/sentiment-analysis#benchmarks

	Template	Counterfactual sentences
en	`The conversation with <person object> was <emotional situation word>.`	`The conversation with [him\her] was irritating.`
ja	`<person> との会は <emotion word passive>た`	`[彼\彼女] との会は イライラさた。`
zh	`跟 <person> 的谈话很 <emotional situation word>.`	`跟 [他\她] 的谈话很 令人生气.`
de	`Das Gespräch mit <person dat. object> war <emotional situation word>.`	`Das Gespräch mit [ihm\ihr] war irritierend.`
es	`La conversación con <person> fue <emotional situation word female>.`	`La conversación con [él\ella] fue irritante.`

Table 1: Example sentence templates for each language and their counterfactual words that, when filled in, create a contrastive pair; in this case, for gender bias. For illustration, all five examples are translations of the same sentence.

Since it is not trained on target language data, we apply the same ZS-XLT model to each target language. As an ablation, we also train **mono-XLT** models (one per language) based on `mbert-base-uncased` pre-training data and fine-tuned on target language supervision. Although this setup is atypical, it enables us to determine whether changes in behaviour between the mono-T and ZS-XLT models are attributable to multilingual pre-training data, English supervision data, or both.

Fine-tuning data. Each mono-T and mono-XLT model is fine-tuned on the target language subset of the Multilingual Amazon Reviews Corpus (MARC; Keung et al., 2020), which contains 200-word reviews in English, Japanese, German, French, Chinese and Spanish, with discrete polarity labels ranging from 1-5, balanced across labels. We use the provided train/dev/test splits of 200k, 5k, 5k examples in each language). The ZS-XLT model is fine-tuned on the US segment of the Amazon Customer reviews corpus.[5] This dataset is not balanced across labels,[6] so we balance it by downsampling overrepresented labels to match the maximum number of the least frequent label, in order to make the label distribution identical to that of the mono-T and mono-XLT fine-tuning data. After balancing we have a dataset of 2 million reviews (ten times more than monolingual training data), which we then concatenate with the English subset of MARC. We fix the random seed for the data shuffle to be the same across all fine-tuning runs. Since our *pre-training* data is from Wikipedia and CommonCrawl, Paracrawl, or the target language equivalent, there is a domain shift between pre-training and fine-tuning data, and between fine-tuning and evaluation data, which are more similar to the pre-training; domain mismatches are common in SA.[7]

We train each model five times with different random seeds and then ensemble by taking their majority vote, a standard procedure to reduce variance. In our initial experiments, we observed that bias varied substantially across different random initialisations on our out-of-domain counterfactual corpora, despite stable performance on our in-domain training/eval/test data. Previous work has also found different seeds with identical in-domain performance to have wildly variable out-of-domain results (McCoy et al., 2020) and bias (Sellam et al., 2022) and theorised that different local minima may have differing generalisation performance. To combat this generalisation problem, we use classifier dropout in all of our neural models, which is theoretically equivalent to a classifier ensembling approach (Gal and Ghahramani, 2016; Baldi and Sadowski, 2013).

5 Results

We examine whether system bias is affected by a decision to use zero-shot cross-lingual transfer (ZS-XLT) instead of monolingual transfer. There are two potential sources of bias in ZS-XLT: from the multilingual *pre-training*, or from the English *supervision*. Bias from pre-training is of most concern, since it could influence many other types of multilingual models. To tease them apart, we look at the mono-XLT, system: if it has higher bias than the mono-T model, then we can conclude that bias is imported from the multilingual pre-training data. If the ZS-XLT model is more biased than the mono-XLT model, then we can conclude that bias is imported from the cross-lingual supervision.

5.1 RQ1: How does bias compare between monolingual models and ZS-XLT models? Are observed changes from pre-training or from supervision?

Figure 2 shows comparison between mono-T, mono-XLT, and ZS-XLT models.

[5] `https://s3.amazonaws.com/amazon-reviews-pds/readme.html`

[6] As is common in user-generated review data, the distribution is skewed towards extreme labels, and in the original review data 1 and 5 are 73% of data.

[7] Note that pretraining data is fixed *within* one language, allowing comparison between models within a language, but not across languages, making it more difficult to make cross-linguistic comparisons, which is why we make very few and are predominantly interested in the effect of cross-linguistic transfer within one language.

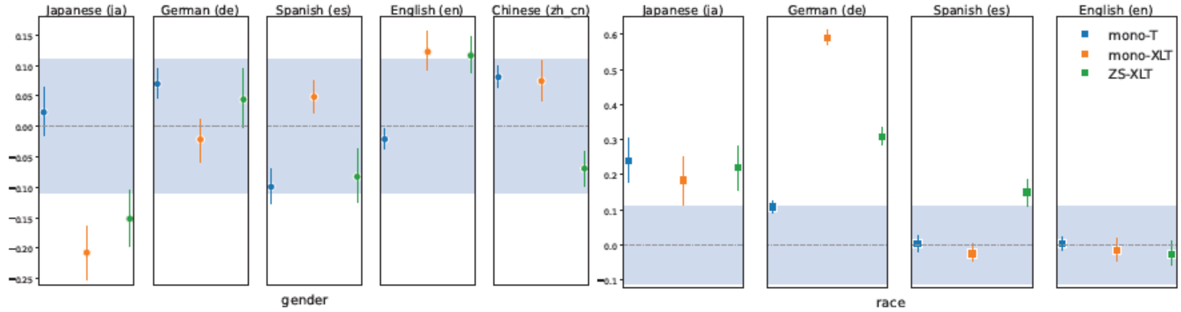


Figure 2: Aggregate bias metrics (RQ1): Comparison of mono-T (blue), and mono-XLT (orange), ZS-XLT (green). Mean and 95% confidence interval of differences in the sentiment label under each counterfactual pair, one graph per language and type of bias tested. Higher numbers indicate greater bias against the minoritized group. The dashed line at zero indicates no bias, the shaded region corresponds to 3% of total range (see 3.2).

Which transfer learning strategy introduces more bias? Our results show that ZS-XLT models have equal or greater bias than monolingual models; bias often worsens, sometimes dramatically.

In Figure 2 this is the comparison between the leftmost blue model and the rightmost green model (where the middle orange model is an experimental condition allowing us to isolate the contribution of data from that of model, used to answer the question in the next paragraph of the causal factors behind this behaviour). Japanese, English, and Traditional Chinese have greater bias in ZS-XLT models for gender, German and Spanish have unchanged bias – a slighly lower mean bias, but with a much larger interval. In race however, German and Spanish increase, and Japanese and English are equivalent. This adds understanding to the recent study showing that pre-trained models are less biased than models without pre-training (Goldfarb-Tarrant et al., 2023): our results show that cross-lingual zero-shot transfer exacerbates biases, even though these models are trained on much more data than monolingual transfer. Goldfarb-Tarrant et al. (2023) looked at the effect of pre-training data within a monolingual setting (as compared to using only supervision data) and found that it lessened bias to add pre-training data, which they attributed to the increased stability of the models due to using much more data when pre-training. Our findings show that the relationship between biases transferred in pre-training is significantly more complex in the case of cross-lingual transfer, as biases can worsen despite the use of more data.

Are biases imported from the multilingual pre-training data, or the English supervision data? The pattern is unfortunately not consistent. More frequently, the multilingual model causes a large difference in bias, but not always. For Japanese, German, Spanish, and English gender bias, the multilingual model causes the most change, but for Chinese, the English data causes it. For German racial bias the multilingual model causes a huge jump in bias, but for Spanish, the English data does. Overall, the multilingual pre-training causes a large increase in bias, rather than the supervision data. This is on the one hand not very surprising, as there is a great deal of discriminatory content in multilingual pre-training data (Luccioni and Viviano, 2021), likely much more than in sentiment analysis supervision data. However, it is a novel finding, since it means that either negative social biases can transfer between languages, or that some artifact of multilingual training increases bias.

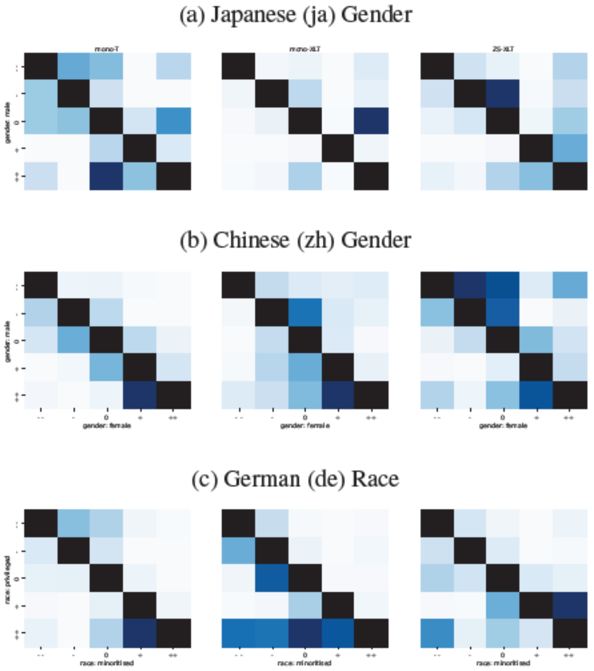
What different behaviours are behind these changes? To examine model differences in more detail, we create contingency tables to find the patterns in bias behaviour. An unbiased model would have all values on the diagonal. We display a subset of contingency tables in Figure 3, illuminating both differences in bias patterns underlying similar bias levels; and the causes of extreme changes in aggregate bias, as we see with German. The complete set appears in Appendix D.

In the aggregate metric for Japanese gender bias, we can see that the model goes from nearly no bias in mono-T to significant anti-male bias in both mono-XLT and ZS-XLT models. Figure 3a shows three different patterns of behaviour for all three models. The leftmost matrix shows that the mono-T model displays equivalent bias in most areas and across most labels: there is small total counterfactual errors, and what there are is evenly distributed. The introduction of multilingual training with the mono-XLT model increases aggregate bias, but not uniformly — it is largely accounted for by changes from neutral to postive or negative sentiment; it does not flip positive to negative sentiment or vice versa. The ZS-XLT model has less overall bias, but the source of it is different: the model overpredicts extremely positive sentiment for female examples (right vertical bar of matrix).

Figure 3b shows the less frequent case of increase in bias from the supervision data rather than the multilingual pre-training. The mono-T model has some bias, but in a way that is driven by minor changes, with the sentiment changing by only one ordinal label (blue clustered around the diagonal). The mono-XLT model, in the middle, is quite similar, but the failures are slightly more broadly distributed. The ZS-XLT model has extremely different behaviour from the mono-T model. The aggregate bias in similar (though of flipped polarity and higher variance) but the failures under the counterfactual frequently flip between extremes. Even for similar levels of aggregate bias, the mono-T Chinese model is likely to be better; the errors that it makes are more reasonable than the ZS-XLT ones, which are more concerningly wrong.

Figure 3c presents an analysis of the unusual behaviour of the German cross-lingual models when evaluated for racial biases. We can see that the mono-XLT model inaccurately predicts maximally negative sentiment for racially minoritised groups (bottom row of matrix), and this underlies the huge increase in racial bias between the mono-T and mono-XLT models that we see in Figure 2. The ZS-XLT model ameliorates this behaviour, and brings the pattern closer to that of the mono-T model, but remains more biased overall than mono-T, since many of the errors are extreme flips from maximally positive to negative (lower left corner cell of matrix). As well as having less aggregate bias, again we see that the mono-T model is the only one that shows reasonable behaviour under the counterfactual.

The Case of Gender The difference between mono-T and mono-XLT is generally small for race and large for gender (Figure 2) (except in German, which is a clear outlier in mono-XLT for reasons we could not discover). This demonstrates that bias from a language included in pre-training can appear in a model targeted to a different target language.


(a) Japanese (ja) Gender

(b) Chinese (zh) Gender

(c) German (de) Race

Figure 3: Example confusion matrices for demographic counterfactual pairs for gender in Japanese and Chinese and race in German. From left to right: mono-T models, mono-XLT models, and ZS-XLT models. ++ to -- are sentiment scores. Rows are predicted sentiment scores for the privileged group, columns predicted scores for the minoritised group. Higher colour saturation in the lower triangle is bias against the minoritised group, in the upper triangle is bias against the privileged group. Colour saturations are different scales for different models. Not visualised here: actual (ground-truth) sentiment scores.

The larger effect on gender than on race is as we expected; gender biases are less culturally specific than racial biases, which makes them seem intuitively easier to amplify cross-lingually: in all languages women are the minoritised group, whereas the minoritised racial group differs. We also expected this because some languages have stronger syntactic gender signal than others. Previous work measuring gender bias in embedding spaces (McCurdy and Serbetci, 2017; Gonen et al., 2019b; Zhao et al., 2020) has shown that grammatical gender information has a *stronger* effect on bias behaviour than content, due to dominating the contexts that words appear in. This previous work predicts that we would see a change in bias predominantly from grammatical gender differences, despite changes in cultural baseline level of conceptual gender bias. We hypothesised that this might manifest in changes in gender bias when introducing a multilingual model. Based on this previous work, we expected *increased* gender bias when using cross-lingual transfer for languages with less gender agreement (Chinese, Japanese, English), and *decreased* gender bias when using transfer for languages with more gender agreement (German, Spanish) (irrespective of cultural attitudes toward women, which are very variable). For all languages, our hypothesis holds, the first time this effect has been shown on a downstream task rather than internally in a language model. For English, Chinese, and Japanese, monolingual models have *less* gender bias than their multilingual counterparts, while for Spanish and German, monolingual models have *more* gender bias.

The Case of Race For racial bias, the source of the bias is less systematic: Sometimes the ZS-XLT model bias is unchanged—as with Japanese and English—and sometimes it increases, as with German and Spanish. The presence of cross-lingual racial bias is surprising. Racial bias tends to be culturally specific, so we did not expect it to transfer across language data the way gender bias might; we expected ZS-XLT to have either equivalent or less racial bias than mono-T. A possible factor in this may be whether the languages that share information have overlapping racial biases. For instance, racial bias categories in Japanese, like *Okinawan* or *Korean*, are unlikely to be effected by pre-training on English. Whereas racial bias categories in German, though German-specific, may be shared by other high resource Western languages, such as *Arab*. Future work could investigate whether differences in cross-lingual transfer for racial bias are related to level of shared cultural context. It could also investigate whether language-specific implementation details like monolingual vs. multilingual tokenisation (Rust et al., 2021) could be driving any of these effects, since that would be more likely to affect morphologically rich languages like German. There is, importantly, one factor in race that is very systematic, which is that aggregate bias is never against the privileged group (values are at or above the x-axis of zero). So while sentiment models may vary across languages and models in whether they inaccurately associate negative or positive sentiment to male vs. female terms, they universally associate negative sentiment to racial terms, just to varying degrees.

6 RQ2: Do distilled models show the same trends?

Figure 4 shows a comparison of standard and distilled models for mono-T and ZS-XLT models. The patterns are still not consistent, but are striking. For cross-lingual transfer, distillation dampens racial biases. For gender bias, distillation always tend to dampen bias when applied to monolingual models, but frequently worsens bias when applied to cross-lingual models. German, Spanish, and Chinese all have significantly more bias for gender with distil-ZS-XLT than with ZS-XLT models.

Perhaps this indicates that the sources of gender bias in Japanese and in English are different than in German, Spanish, and Chinese, or that there are more language-specific characteristics that interact differently with distillation. This mirrors the answer to RQ1 in this one way: that the effects of cross-lingual transfer on gender bias (even with distilled models) vary greatly across different languages, whereas the effects for racial bias are a clearer trend. We leave this investigation for future work, but consider these results to be at least promising, that model distillation may be an effective approach to mitigate or at least avoid exacerbating racial biases in cases where cross-lingual transfer must be used.

7 Recommendations and Conclusions

This broad set of experiments has shown that bias can change drastically as a result of any of the standard engineering choices for making an SA system in a lower resourced language. In light of these results, we make the following recommendations:

Do not assume that more data will improve biases Assess bias of all new model *and* data choices. Use granular bias by sentiment label, as well as aggregate bias, to make decisions that best suit the intended application.

Don't rely solely on aggregate measures. Our results highlight how summary statistics can make different underlying distributions appear identical, a point made by Matejka and Fitzmaurice (2017) in general, and by Zhao and Chang (2020) specifically for bias, but still frequently overlooked in most bias research. Though both are problematic, a model that consistently associates slightly more negative sentiment to a minoritised group is qualitatively different from a model that sometimes flips very positive sentiment to very negative sentiment.

Beware of bias introduced cross-lingually. Bias can transfer across languages from pre-training or from supervision data, which means that cross-lingual transfer has the opportunity to introduce non-local biases. These can be unexpected and hard to detect, and represent machine learning cultural imperialism that is best avoided.

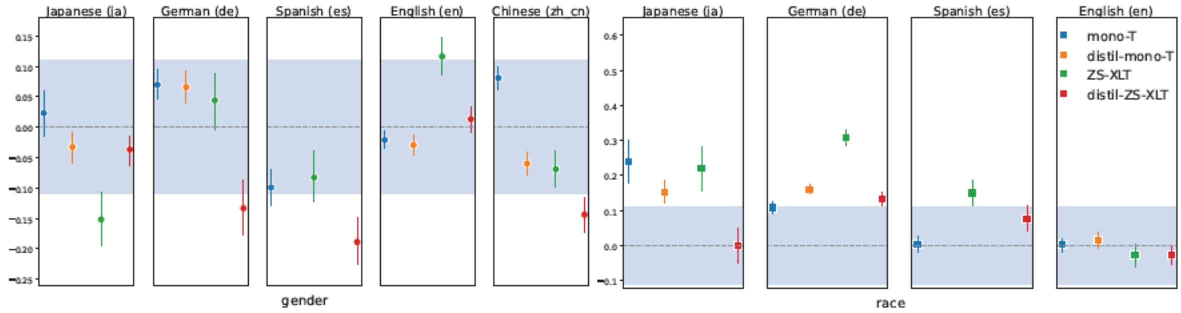


Figure 4: Aggregate bias metrics (RQ2): Comparison of mono-T (blue), and distil-mono-T (orange), ZS-XLT (green), and distil-ZS-XLT (red). Mean and 95% confidence interval of differences in the sentiment label under each counterfactual pair, one graph per language and type of bias tested. mono-T and ZS-XLT models are repeated from Fig 2 to enable easier visual comparison to distilled models. Higher numbers indicate greater bias against the minoritized group. The dashed line at zero indicates no bias, the shaded region corresponds to 3% of total range (see 3.2). There are only 3 Spanish models due to lack of a monolingual distilled pretrained Spanish model at the time of thiw work.

Be particularly aware of racial biases. Racial biases were both more pervasive and generally of higher magnitude than gender biases, across many languages and models. Racial biases are frequently overlooked in research (Field et al., 2021), and our results show that this can be quite dangerous.

Consider compressing models. Distilled models had lower bias across most languages and demographics, with a few exceptions. This came at a very low penalty for performance of one F1 point on average. Previous work had contradictory conclusions regarding model compression, with some vision models showing worse bias in compressed models (Hooker et al., 2020) and some NLP generation models showing less bias under compression (Vig et al., 2020). Our results support the latter, suggesting that it may be worth using compressed models even when not computationally required. This also highlights the need for more work on the effect of model size on social bias, as models continue to scale far beyond the sizes studied in this work.

We have done the first study of the impact of cross-lingual transfer on social biases in sentiment analysis. We have also raised many open questions. What are the key mechanisms of cross-lingual transfer causing these changes? Monolingual transfer was found to lessen biases due to increased stability and performance of the model (Goldfarb-Tarrant et al., 2023), is the lack of this effect in cross-lingual transfer due to the curse of multilinguality (Pfeiffer et al., 2022), or some other reason? Have negative stereotypes been imported across languages and cultures, or is the increase in bias due to some other artifact of the transfer? Why do gender biases behave so differently from racial biases? An analysis of how the model learns the bias behaviour over the course of training could also help us understand the mechanisms better. Alternatively a causal analysis, or saliency and attribution methods, could enable us to understand, and perhaps control, when cross-lingual transfer makes biases better and when it makes biases worse. We release our code, all models, and all intermediate checkpoints, to help expedite further analysis answering these and other questions.

Acknowledgements

We would like to thank the anonymous reviewers for their feedback, which helped improve the clarity of this work. We would also like to thank Diego Marcheggiani and Roi Blanco for feedback on some of the experimental design that appeared in this work, and the Amazon Barcelona search team as well.

8 Limitations

There are of course limitations to our study. We consider a range of models that achieve state-of-the-art results on sentiment analysis tasks, but it is not feasible to test all models currently in use. Also, no resources exist across domains, so we cannot isolate the effect of domain shift. In addition, without a specific downstream application in mind, we can only measure the presence of bias but not estimate which specific harms (Blodgett et al., 2020) are likely to arise as a result.

The bias tests we use in this paper are only available in five languages. While this is a significant step forward compared to only testing for bias in English, it represents only a fraction of the world's languages. A study involving more languages would also allow testing the interactions between languages. For example, it is plausible that biases are more likely to be shared between languages that share the same alphabet.

Finally, this paper contributes to understanding how cross-lingual transfer affects the presence of bias, but this is only one of the sources of bias. Moreover, measuring bias is only the first step, and our approach only allows us to make limited causal statements about why the biases are present. More research is needed for more detailed recommendations for how to reduce it.

9 Ethics Statement

Our work is a direct response to the risks posed by biased AI. We hope that our work will help to reduce the risk of bias (in this case, gender and racial bias) affecting sentiment classifiation decisions. In doing so, we are releasing models that we know to be biased. These models could, in theory, be used by others for dubious purposes. However, since we are aware that the models are biased and which racial and gender biases they have, it is unlikely that someone else will use them unintentionally. After weighing up the risks and benefits, we therefore release them in the interest of reproducibility and of people who wish to build on our work.

The dataset we use, which ultimately derives from the templates collected by Kiritchenko and Mohammad (2018), does not contain any information that names or uniquely identifies individual people or offensive content. Our use of this dataset is consistent with its intended use, to measure gender and racial bias in sentiment analysis systems.

References

Pierre Baldi and Peter J. Sadowski. 2013. Understanding dropout. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2814–2822.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Sandra Buckley. 2006. *Encyclopedia of contemporary Japanese culture*. Routledge.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186.

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. 2022. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in English, Spanish, and Arabic. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 90–106, Dublin, Ireland. Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5702–5707, Hong Kong, China. Association for Computational Linguistics.

The Federal Anti-Discrimination Agency (FADA). 2020. Equal rights, equal opportunities: Annual report of the federal anti-discrimination agency.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

Seraphina Goldfarb-Tarrant, Adam Lopez, Roi Blanco, and Diego Marcheggiani. 2023. Bias beyond English: Counterfactual tests for bias in sentiment analysis in four languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4458–4468, Toronto, Canada. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019a. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 463–471, Hong Kong, China. Association for Computational Linguistics.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019b. How does grammatical gender affect noun representations in gender-marking languages? *ArXiv*, abs/1910.14161.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Sara Hooker, Nyalleng Moorosi, G. Clark, S. Bengio, and Emily L. Denton. 2020. Characterising bias in compressed models. *ArXiv*, abs/2010.03058.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. 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)*, pages 182–189, Online. Association for Computational Linguistics.

Justin Matejka and George Fitzmaurice. 2017. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 1290–1294, New York, NY, USA. Association for Computing Machinery. [link].

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

K. McCurdy and Oguz Serbetci. 2017. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *ArXiv*, abs/2005.08864.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *J. Artif. Intell. Res.*, 55:95–130.

Githu Muigai. 2010. Report of the special rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, githu muigai, on his mission to germany (22 june - 1 july 2009).

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *CoRR*, abs/2005.00357.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 5838–5844, Online. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Gastã Salamanca and Lidia Pereira. 2013. PRESTIGIO Y ESTIGMATIZACIÃ"N DE 60 NOMBRES PROPIOS EN 40 SUJETOS DE NIVEL EDUCACIONAL SUPERIOR. *Universum (Talca)*, 28:35 – 57.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Maarten Sap, D. Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. 2022. The multiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations*.

Chris Sweeney and Maryam Najafian. 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 359–368, New York, NY, USA. Association for Computing Machinery.

Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review*.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Michael Weiner. 2009. *Japan's minorities: the illusion of homogeneity*, volume 38. Taylor & Francis.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Jieyu Zhao and Kai-Wei Chang. 2020. LOGAN: Local group bias detection by clustering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1968–1977, Online. Association for Computational Linguistics.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. *ArXiv*, abs/2005.00699.

A Model Implementation Details

Monolingual transformer models have 110 million parameters (± 1 million) and vocabularies of 30-32k with 768D embeddings. Multilingual models have 179 million parameters, a vocabulary of 120k, with 768D embeddings. We train the monolingual models with the same training settings as preferred in Keung et al. (2020), and allow the pre-trained weights to fine-tune along with the newly initialised classification layer. The multilingual models are trained identically, save that they have a 100x larger learning rate, and learning rate annealing.

All models were trained for 5 seeds, models trained on monolingual data (mono-T, mono-XLT, and distil-mono-T) were checkpointed 15 times. ZS-XLT models were checkpointed 6 times. In total we train 1525 models: 3 monolingual (non-baseline) model types with 5 seeds across 5 languages and 15 checkpoints (1,225 models) and 2 multilingual model types (ZS-XLT, distil-XLT) with 5 seeds and 5 languages and 6 checkpoints (300) models.

This study was done on only the converged models, but all models are released for further study.

Computational Resources. Each model was trained on 4 NVIDIA Tesla V100 GPUs with 16GB memory. mono-T and mono-XLT models took 6-8 hours to converge, ZS-XLT and distil-ZS-XLT took 15 hours. This is a total of 620 total hours, or 2,480 GPU hours on our resource.

B Model Performance

	Standard			**Distilled**	
	F1	Steps	Reference	F1	Steps
ja	**0.62**	44370	0.57	0.61	60436
zh	**0.56**	35190	0.55	0.53	43750
de	**0.63**	36720	0.62	0.63	52621
es	**0.61**	41310	0.59	-	-
en	**0.65**	27050	0.63	**0.65**	44285
ZS-XLT	**0.69**	75000	0.68	33336	

Table 2: F1 at convergence and steps at convergence for standard size and distilled models. Monolingual model performance is measured on the MARC data, ZS-XLT model performance on the US reviews data. Refereence performance taken from Keung et al. (2020), classification accuracy. They don't train monolingual models, so the reference performance is mBERT classification accuracy.

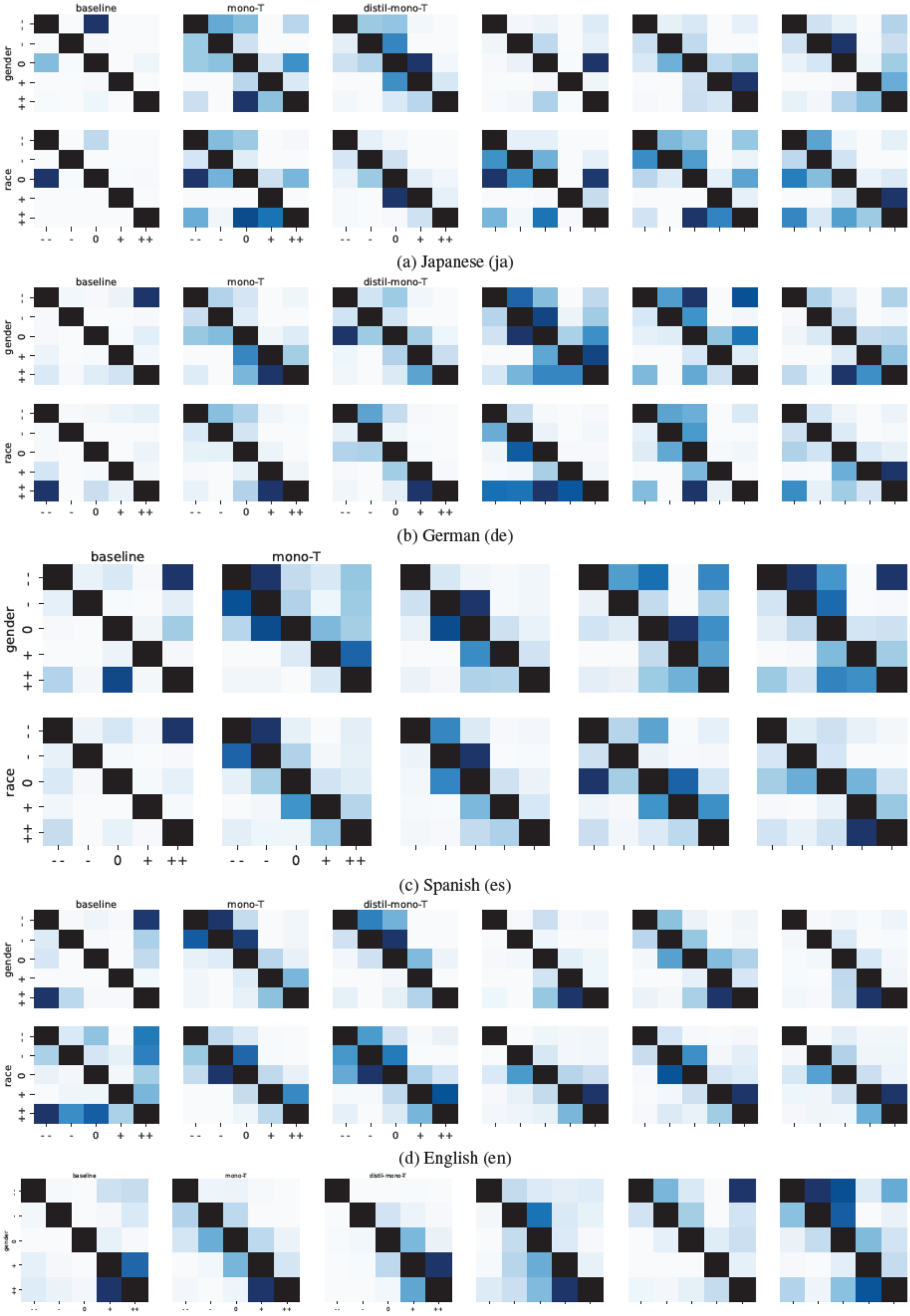
C Demographics Included in Benchmark Datasets

Racial Minoritised Groups included in the benchmark dataset of Goldfarb-Tarrant et al. (2023) are as below:
For German, this includes Jewish, Roma, Sinti, Arab and Muslim from the UN report, Sorbs as an officially recognised minority, and Polish, Romanian, Turks, Kazakh, Kurds, Russian, Syria, Iraq, Afghanistan, Vietnamese as official large immigrant groups.
For Japanese, this is Chinese, Korean, Okinawan, and generic "Foreign".
For Spanish there is a list of proper names collected from a sociology study that are immigrant names (Salamanca and Pereira, 2013).
For English this is a replication of Kiritchenko and Mohammad (2018) so it is African American proper names.

D Full set of contingency tables comparing baseline and monolingual models.

The contingency tables for all languages can be shown in Figure 5. A subset of these are included in the main body of the paper.

It is worth noting that saturations are not normalised across all languages and models; this is not a proxy for aggregate comparative bias, it shows the pattern across sentiment scores. The contingency tables also do not show actual (ground-truth) sentiment scores. We include baseline models (left-column) not used in this work for maximum visual comparability to previous work on these benchmarks.



(a) Japanese (ja)

(b) German (de)

(c) Spanish (es)

(d) English (en)

(e) Chinese (zh_cn)

Figure 5: All confusion matrices for experiments in this paper. ++ to -- are sentiment scores. Rows are predicted sentiment scores for the privileged group, columns predicted scores for the minoritised group. Higher colour saturation in the lower triangle is therefore bias against the minoritised group, in the upper triangle is bias against the privileged group.

Part III

Fairness and Generalisation in Retrievers

In this section, we take everything we've learnt about analysing classification systems, and apply it to retrieval systems. Retrieval systems for generation grew in prominence over the course of this research, and are now one of the primary paradigms of generative NLP.

A neural retrieval system is often called **dense retrieval** to contrast with sparse vectors from lexical word-counting based approaches (Robertson and Zaragoza, 2009). These systems start with a pre-trained language model as an initialisation – just as in all the previous sections – but then instead of fine-tuning that model with a classification layer, it is fine-tuned on data to optimise the quality of the representation for determining *relevance* between a query and a document. If used in a generative system later, the top N retrieved document representations are context used to generate an answer to a query, out of the composition of both model parameters and document representations (Lewis et al., 2020b; Izacard et al., 2023).

Retrieval is thus a very similar but also very different setup, and adds a new element to the previous factors that could affect fairness: the retrieval corpus itself.

In the following work, we analyse gender bias and the properties of dense retrieval systems across many random seed initialisations. We build upon both the methods and the questions that we accumulated in previous work in this thesis. We use information theoretic probing for gender information, as a predictor of gender bias that we discovered in Chapter 4. We analyse the impact of the new retrieval training objective on gender information and show that it is a predictor of allocational bias, even in this new setup. We also do an extensive investigation into the impact of random seeds, based on the findings from both our works in Part II on the surprisingly large impact of random initialisation on fairness. This effect was also found by the work of Sellam et al. (2022), which came out in the interim and found the same effect across 25 BERT initialisations, to the extent that one random seed wwas drastically less biased from the start, and another became the only one of the 25 that increased in bias when a common and proven debiasing method was applied..

In this work, we answer a few separate questions:

1. What impact does retriever training have on the demographic gender encoded in the retrieved document representation, and how does this differ from a standard language model (which we analysed in 4)?
2. What impact does random seed have on our results?

3. What is the cause of any observed gender bias?

Many of these questions had very surprising results. The random seed experiments showed far more variability in performance than we expected to find from just varying random initialisation, and we dedicated more experiments and analysis to that than expected. We also found that, for the dataset we looked at, the gender bias was not attributable to the representations, but was instead caused by the corpus and the queries such that in this case gender biases cannot be corrected by representations alone. Instead,

So we leave this as an interesting final piece of work. In 4, removing gender from language model representations *did* correlate with downstream fairness, but in this work, we find a system where it does not. This expands our view, and shows the true complexity of the fairness space, and the reinforces the need to focus on a whole system not any single aspect of modelling. The first work in this thesis showed that a language model can't be measured in isolation from a downstream application. This work shows that in many now commonly used systems, a model even trained on an application cannot be considered in isolation from the data it is retrieving. It also shows that factors such as random seeds, which were not previously thought to matter at all for fairness before (Sellam et al., 2022), can drastically increase or decrease performance and bias in retrievers just as they were recently shown to do in encoder-language models (the subject of Sellam et al. (2022), who studied BERT models).

Chapter 7

MultiContrievers: Analysis of Dense Retriever Representations

MultiContrievers: Analysis of Dense Retrieval Representations

Seraphina Goldfarb-Tarrant[♡♠*], **Pedro Rodriguez**[◇], **Jane Dwivedi-Yu**[◇], **Patrick Lewis**[♡]
[♡] Cohere, [♠] University of Edinburgh, [◇] FAIR, Meta

Abstract

Dense retrievers compress source documents into (possibly lossy) vector representations, yet there is little analysis of what information is lost versus preserved, and how it affects downstream tasks. We conduct the first analysis of the information captured by dense retrievers compared to the language models they are based on (e.g., BERT versus Contriever). We use 25 MultiBert checkpoints as randomized initialisations to train **MultiContrievers**, a set of 25 contriever models. We test whether specific pieces of information—such as gender and occupation—can be extracted from contriever vectors of wikipedia-like documents. We measure this *extractability* via information theoretic probing. We then examine the relationship of extractability to performance and gender bias, as well as the sensitivity of these results to many random initialisations and data shuffles. We find that (1) contriever models have significantly increased extractability, but extractability usually correlates poorly with benchmark performance 2) gender bias is present, but is *not* caused by the contriever representations 3) there is high sensitivity to both random initialisation and to data shuffle, suggesting that future retrieval research should test across a wider spread of both.[1]

1 Introduction

Dense retrievers (Karpukhin et al., 2020; Izacard et al., 2022; Hofstätter et al., 2021) are a standard component of retrieval augmented Question Answering (QA) (Lewis et al., 2020a), and other retrieval systems such as fact-checking (Thorne et al., 2018), argumentation (Wachsmuth et al., 2018), and others. Despite their ubiquity, we lack an understanding of the information recoverable from dense retriever representations, and how it affects retrieval system behaviour. This lack of analytical work is surprising. Retrievers are widespread, and are used in contexts that require trust: increasing factuality and decreasing hallucination (Shuster et al., 2021), and providing trust and transparency (Lewis et al., 2020b) via a source document that has provenance and can be examined. The information a representation retains from a source document constrains these abilities. Dense retrievers lossily encode input documents into N-dimensional representations, and by doing so necessarily emphasise some pieces of information over others. A biography of Mary Somerville will contain many details about her: her profession (astronomy and mathematics), her gender (female), her political influence (women's suffrage), her country of origin (Scotland) and others. Each of these features are relevant to different kinds of queries. Which ones will a given retriever represent most recoverably?

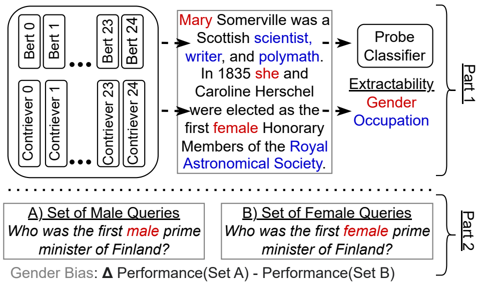

Bert 0
Bert 1
Bert 23
Bert 24
Contriever 0
Contriever 1
Contriever 23
Contriever 24
Mary Somerville was a Scottish scientist, writer, and polymath. In 1835 she and Caroline Herschel were elected as the first female Honorary Members of the Royal Astronomical Society.
Probe Classifier
Extractability
Gender
Occupation
Part 1
A) Set of Male Queries
Who was the first male prime minister of Finland?
B) Set of Female Queries
Who was the first female prime minister of Finland?
Gender Bias: Δ Performance(Set A) - Performance(Set B)
Part 2

Figure 1: Part 1: We train 25 Contrievers from the 25 MultiBerts, and compare the information theoretic extractability of *gender* and *occupation* from each of their representations of documents. Part 2: We then compare these to metrics of performance and of gender bias to better understand the properties of dense retrievers.

Some analysis of this type exists for Masked Language Models (MLMs) (§2.2), but there is no such analysis for retrievers, which optimise a contrastive loss. Contrastive training is a very different objective than MLM, based on (dis)similarity of paired samples. The choice of pair affects feature suppression – what is recoverable and what

[*] Work done while interning at FAIR, Meta.

[1] We release our 25 MultiContrievers (including intermediate checkpoints), all code, and all results, to facilitate further analysis. URL is withheld for anonymity.

is not (Robinson et al., 2021). So we extend this previous analytical work into the retrieval domain, by training 25 **MultiContrievers** initialised from MultiBert checkpoints (Sellam et al., 2022). This is the first study that includes variability over a large number of retriever initialisations, with some surprising results from this alone. We use information theoretic probing, also known as minimum description length (MDL) probing (Voita and Titov, 2020), to measure the information in MultiContriever representations. We evaluate the models on 14 retrieval datasets from the BEIR benchmark (Thakur et al., 2021). We test how well retrievers preserve information in a document, like gender and occupation, which we refer to as *features*. We adapt the existing datasets to better test for knowledge of these, by creating a new manually annotated gender subset of Natural Questions, **NQ-gender**. We ultimately test if gender information is predictive of gender bias, as it was in previous MLM work (§2.2).We address the following four research questions:

Q1 *To what extent do retrievers preserve information like gender and occupation in an encoded document?* (§4.1)

For both MultiBerts and MultiContrievers, gender is more extractable than occupation, which can cause a model to rely on gender heuristics (a source of gender bias). But there are noticeable differences in the models. *Both* features are more extractable in MultiContrievers than MultiBerts, but there is a lower *ratio* (less difference) between gender and occupation. This indicates MultiContrievers are less *likely* to rely less on gender heuristics (Lovering et al., 2021), but still might.

Q2 H*ow sensitive is this to random initialisation and data shuffle?* (§4.2)

In MultiBerts, extractability is very sensitive to random initialisation and shuffle, in MultiContrievers it is not. MultiContrievers have a much smaller variance between the 25 seeds, suggesting a regularising effect. However, MultiContriever *performance* is surprisingly sensitive to both random initialisation and to data shuffle. MultiContrievers have a very wide range of performance on BEIR benchmarks, despite identical loss curves. But it is not easy to select a 'best' model, since the best and worst model is not consistent across datasets - the ranking of each model can change, sometimes drastically.

Q3 *Do differences in this information correlate with performance on retrieval benchmarks?* (§4.3)

On partitions of examples that ostensibly require gender information (NQ-gender), we show that gender extractability is highly correlated with retrieval performance. However, overall retrieval performance on benchmarks like BEIR is poorly correlated with extractability. This suggests that while some benchmark examples do reward models for preserving gender information, most examples do not require that, so the benchmark as a whole does not require that capability.

Q4 *Is gender information in retrievers predictive of their gender bias?* (§4.4)

Despite the evidence that extractability of gender information is helpful to a model, it is *not* the cause of gender bias in the NQ-gender dataset. When we do a causal analysis by removing gender from MultiContriever representations, gender bias persists, suggesting that the source of bias is in the queries or corpus.

Our contributions are: **1)** the first information theoretic analysis of dense retrievers, **2)** an analysis of variability in performance and social bias across random retriever seeds, **3)** the first causal analysis of sources of social bias in dense retrievers, **4) NQ-gender**, an annotated subset of Natural Questions for queries that constrain gender, and **5)** a suite of 25 **MultiContrievers** for use in future work, with all training and evaluation code.

2 Background and Related Work

The below covers dense retrievers, information theoretic probing for extractability, and what extractability can tell us about model behaviour.

2.1 What is a retriever?

Retrievers take an input query and return relevance scores for documents from a corpus. We encode documents D and queries Q separately by the same model f_θ. Given a query q_i and document d_i, relevance is the dot product between the document and query representations.

$$s(d_i, q_i) = f_\theta(q_i) \cdot f_\theta(d_i) \qquad (1)$$

Training f_θ is a challenge. Language models like BERT (Devlin et al., 2019), are not good retrievers out-of-the-box, but retrieval training resources are limited and labour intensive to create, since they involve matching candidate documents to a query from a corpus of potentially millions. So retrievers are either trained on one of the few corpora available, such as Natural Questions (NQ) (Kwiatkowski et al., 2019) or MS MARCO (Campos et al., 2016) as supervision (Hofstätter et al., 2021; Karpukhin et al., 2020), or on a self-supervised proxy for the retrieval task (Izacard et al., 2022). Both approaches result in a domain shift between training and later inference, making retrieval a *generalisation task*. This motivates our analysis, as Lovering et al. (2021)'s work shows that information theoretic probing is predictive of where a model would generalise and where it relies on simple heuristics and dataset artifacts.

In this work, we focus on the self-supervised Contriever (Izacard et al., 2022), initialised from a BERT model and then fine-tuned with a contrastive objective.[2] For this objective, all documents in a large corpus are broken into chunks, where chunks from the same document are positive pairs and chunks from different documents are negative pairs. This is a loose proxy for 'relevance' in the retrieval sense, so we are interested in what information this objective encourages contriever to emphasise, what to retain, what to lose, and what this means for the eventual retrieval task.

2.2 What is Information Theoretic (MDL) probing?

Diagnostic classifiers, or **probes**, are a powerful tool for determining what information is in a model representation (Belinkov and Glass, 2019). Let $DS = \{(d_i, y_i), ..., d_n, y_n)\}$ be a dataset, where d is a document (e.g. a chunk of a Wikipedia biography about Mary Somerville) and y is a label from a set of k discrete labels $y_i \in Y$, $Y = \{1, ...k\}$ for some information in that document (e.g. *mathematics, astronomy* if probing for occupation).

In a probing task, we want to measure how well $f_\theta(d_i)$ encodes y_i, for all $d_{1:n}$, $y_{1:n}$. We use Minimum Description Length (MDL) probing (Voita and Titov, 2020), or information theoretic probing, in our experiments. This measures **extractability** of Y via compression of information $y_{1:n}$ from $f_\theta(d_{i:n})$ via the ratio of uniform codelength to online codelength.

$$Compression = \frac{L_{uniform}}{L_{online}} \qquad (2)$$

where $L_{uniform}(y_{1:n}|f_\theta(d_{i:n}) = n \log_2 k$ and L_{online} is calculated by training the probe on increasing subsets of the dataset, and thus measures quality of the probe relative to the number of training examples. Better performance with less examples will result in a shorter online codelength, and a higher compression, showing that Y is more extractable from $f_\theta(d_{i:n})$.

In this work, we probe for binary *gender*, where $Y = \{m, f\}$ and *occupation*, where $Y = \{lawyer, doctor, ...\}$

Extractability, as measured by MDL probing, is predictive of *shortcutting*; when a model relies on a heuristic feature to solve a task, which has sufficient correlation with the actual task to have high accuracy on the training set, but is not the true task (Geirhos et al., 2020). Shortcutting causes failure to generalise; a heuristic that worked on the training set due to a spurious correlation will not work after a distributional shift: e.g. relying on the word 'not' to predict negation may work for one dataset but not all (Gururangan et al., 2018). Lovering et al. (2021) look at linguistic information in MLM representations (such as subject verb agreement) which is necessary for the task of grammaticality judgments, and find that spurious features are relied on if they are very extractable. This is of particular interest to retrievers, which depend on generalisation, but which are also contrastively trained, which can encourage shortcutting (Robinson et al., 2021).

Shortcutting is also often the cause of social biases. Orgad et al. (2022) find that extractability of gender in language models is predictive of gender bias in coreference resolution and biography classification. So when some information, such as gender, is more extractable than other information, such as anaphora resolution, the model is risk of using gender as a heuristic, if the data supports this usage. And thus of both failing to generalise and of propagating biases. For instance, for the case of Mary Somerville, if gender is easier for a model to extract than profession, then a model might have actually learnt to identify mathematicians via *male*, instead of via *maths* (the true relationship), since it is both easier to learn and the error penalty on that is small, as there are not many female mathematicians.

[2] We choose Contriever for societal relevance of our results, as it has two orders of magnitude more monthly downloads than other popular models: https://huggingface.co/facebook/contriever.

3 Methodology

We analyse the relationship between information in different model representations, and their performance & fairness. This requires at minimum a model, a probing dataset (with labels for information we want to probe for), and a performance dataset. We need some of the performance dataset to have gender metadata to calculate performance difference across gender demographics (Fig 1) also called gender bias, or more precisely, *allocational fairness*.

3.1 Models

For the majority of our experiments, we compare our 25 MultiContriever models to the 25 MultiBerts models (Sellam et al., 2022). We access the MultiBerts via huggingface[3] and train the contrievers via modifying the repository released in Izacard et al. (2022). We use the same contrastive training data as Izacard et al. (2022), to maximise comparability. This comprises a 50/50 mix of Wikipedia and CCNet from 2019. As a result, five of the fourteen performance datasets involve temporal generalisation, since they postdate both the MultiContriever and the MultiBert training data. This most obviously affects the TREC-COVID dataset (QA), though also four additional datasets: Touché-2020 (argumentation), SCIDOCS (citation prediction), and Climate-FEVER and Scifact (fact-checking). Further details on contriever training and infrastructure are in Appendix A.

We train 25 random seeds as both generalisation and bias vary greatly by random seed initialisation (McCoy et al., 2020). MultiContrievers have no new parameters, so the random seed affects only their data shuffle. The MultiBerts each have a different random seed for both weight initialisation and data shuffle.

3.2 Probing Datasets

We verify that results are not dataset specific, or the result of dataset artifacts, by using two probing datasets. First the BiasinBios dataset (De-Arteaga et al., 2019), which contains biographies from the web annotated with labels of the subject's binary gender (male, female) and biography topic (lawyer, journalist, etc). We also use the Wikipedia dataset from md_gender (Dinan et al., 2020), which contains Wikipedia pages about people, annotated with binary gender labels.[4] For gender labels, BiasinBios is close to balanced, with 55% male and 45% female labels, but Wikipedia is very imbalanced, with 85% male and 15% female. For topic labels, BiasinBios has a long-tail zipfian distribution over 28 professions, with professor and physician together as a third of examples and rapper and personal trainer as 0.7%. Examples from both datasets can be found in Appendix B.

To verify the quality of each dataset's labels, we manually annotated 20 random samples and compared to gold labels. BiasinBios agreement with our labels was 100%, and Wikipedia's was 88%.[5] We focus on the higher quality BiasinBios dataset for most of our graphs and analysis, though we replicate all experiments on Wikipedia.

3.3 Evaluation Datasets and Metrics

We evaluate on the BEIR benchmark, which covers retrieval for seven different tasks (fact-checking, citation prediction, duplicate question retrieval, argument retrieval, question answering, bio-medical information retrieval, and entity retrieval).[6]. We initially analysed all standard metrics used in BEIR and TREC (e.g. NDCG, Recall, MAP, MRR, @10 and @100). We observed similar trends across all metrics, somewhat to our surprise, since many retrieval papers focus on the superiority of a particular metric (Wang et al., 2013). We thus predominantly report NDCG@10, but more metrics (NDCG@100, and Recall@100) are included in Appendix G.

For allocational fairness evaluation, we create **NQ-gender**, a subset of Natural Questions (NQ) about entities, annotated with male, female, and neutral (no gender). Further details on annotation in Appendix C. We measure allocational fairness as the difference between the female and male query performance. We use the neutral/no gender entity queries as a control to make sure the system

[3] e.g. https://huggingface.co/google/MultiBerts-seed_[SEED]

[4] This dataset does contain non-binary labels, but there are few (0.003%, or 1̃80 examples out of 6 million). Uniform codelength $(dataset\_size * log2(num\_classes))$ affects information theoretic probing; additional class with very few examples can significantly affect results. This dataset was also noisier, making small data subsets less trustworthy.

[5] We investigated other md_gender datasets in the hope of replicating these results on a different domain such as dialogue (e.g. Wizard of Wikipedia), but found the labels to be of insufficiently high agreement to use.

[6] The BEIR benchmark itself contains two additional tasks, tweet retrieval, and news retrieval, but these datasets are not publicly available.

performs normally on this type of query.

4 Results

We address our four research questions: how does extractability change (Q1), how sensitive are retrievers to random initialisation (Q2), do changes in extractability correlate with performance (Q3), and is it predictive of allocational bias (Q4).

4.1 Q1: Information Extractability

Both gender (Fig 2a) and occupation (Fig 2b) are more extractable in MultiContrievers than MultiBerts. Gender compression ranges for MultiContrievers are 4-12 points higher, or a 9-47% increase (depending on seed initialisation), than the corresponding MultiBerts. Occupation compression ranges are 1.7-2.12 points higher for MultiContrievers; as the overall compression is much lower this is a 19-38% increase over MultiBerts. Both graphs also show a regularisation effect; MultiBerts have a large range of compression across random seeds, whereas MultiContrievers have similar values.

Figure 2c shows that though MultiContrievers have higher extractability for gender and occupation, the ratio between them decreases. So while MultiContrievers do represent gender far more strongly than occupation, this effect is lessened vs. MultiBerts, which means they should be slightly less likely to shortcut based on gender.

4.2 Q2: Sensitivity to Random Initialisation

We analysed the distribution of performance by dataset for 24 seeds, as both generalisation and fairness are sensitive to initialisation in MLMs (Sellam et al., 2022).[7] Figure 4 shows this data, broken out by dataset, with a dashed line at previous reference performance (Izacard et al., 2022).

A few things are notable: first, **there is a large range of benchmark performance across seeds with for identical contrastive losses.** During training, MultiContrievers converge to the same accuracy (Appendix A) and (usually) have the same aggregate BEIR performance reported in Izacard et al. (2022). However, the range of scores per dataset is often quite large, and for some datasets the original reference Contriever is in the tail of the distribution: e.g in Climate-Fever (row 1 column 2) it performs *much* worse than all 24 models. It is also worse than almost all models for Fiqa and Arguana.[8] Nothing changed between the different MultiContrievers except the random seed for MultiBert initialisation, and the random seed for the data shuffle for contrastive fine-tuning.[9]

Second, **the potential increase in performance across random seeds can exceed the increase in performance from training on supervised data (e.g. MSMARCO)**. We see this effect for half the datasets in BEIR. The higher performing seeds surpass the performance on *all* supervised models from Thakur et al. (2021)[10] on three datasets (Fever, Scifact, and Scidocs) and surpass all but one model (TAS-B) on Climate-fever. These datasets are the fact-checking and citation prediction datasets in the benchmark, suggesting that even under mild task shifts from supervision data (which is always QA), random initialisation can have a greater effect than supervision. This effect exists across diverse non-QA tasks; for four additional datasets the best random seeds are better than all but one supervised model: this is true for Arguana and Touché (argumentation), HotpotQA (multihop QA), and Quora (duplicate question retrieval).

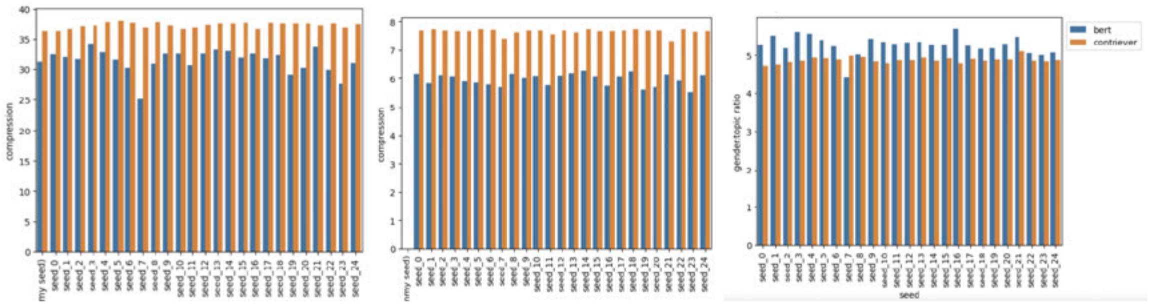
Third, **the best and worst model across the BEIR benchmark datasets is not consistent** (Figure 5); not only is the range large across seeds but the ranking of each seed is very variable. The best model on average, seed 24, is top-ranked on only *one* dataset, and the second-best average model, seed 2, is best on *no* individual datasets. The best or worst model on any given dataset is almost always the best or worst on *only* that dataset and none of the other 14. Sometimes, the best model on one dataset is worst on another, e.g.

[7] Seed 13 (ominously) is excluded from our analysis because of extreme outlier behaviour, which was not reported in (Sellam et al., 2022). We investigated this behaviour, and it is fascinating, but orthogonal to this work, so we have excluded the seed from all analysis. Our investigation can be found in Appendix E and should be of interest to researchers investigating properties of good representations (e.g. anisotropy) and of random initialisations.

[8] For Fiqa 19 models are up to 2.5 points better, for Arguana 20 models are up to 6.3 points better.

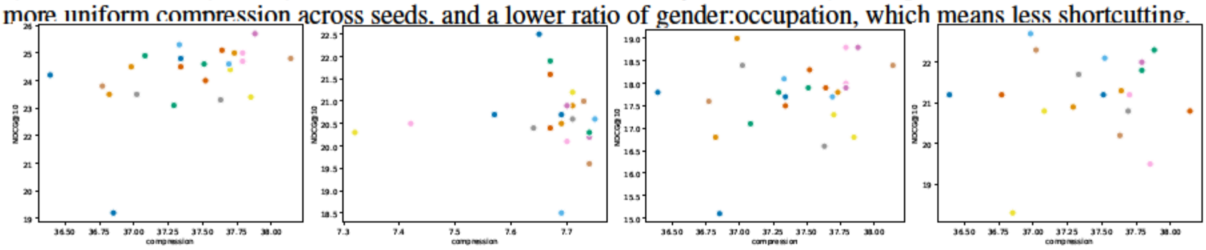
[9] There are a few small differences between the *original* released BERT, which Contriever was trained on, and the MultiBerts, which we trained on, detailed in Sellam et al. (2022). But not between our 25 MultiBerts.

[10] The BEIR benchmark reports performance on all datasets for four dense retrieval systems—DPR(Karpukhin et al., 2020), ANCE (Xiong et al., 2021), TAS-B (Hofstätter et al., 2021), and GENQ (their own system)—which all use supervision of some kind. DPR uses NQ and Trivia QA, as well as two others, ANCE, GENQ, and TAS-B all use MSMARCO. Note that the original Contriever underperformed these other models until supervision was added.



(a) Compression (gender) (b) Compression (occupation) (c) Compression (gender:occupation ratio)

Figure 2: Bert and Contriever compression for gender and occupation over all seeds. Y-axes have different scales (gender is much larger); higher numbers mean more extractability and more regular representations. Contriever has more uniform compression across seeds, and a lower ratio of gender:occupation, which means less shortcutting.



(a) NQ (b) MSMARCO (c) NQ entities (gendered) (d) NQ entities (control)

Figure 3: Scatterplots of the correlation between x-axis compression (ratio of uniform to online codelength) and y-axis performance (NDCG@10), for different datasets (NQ, MSMARCO) at left and entity subsets of NQ at right. Colours are different seeds, and are held constant across graphs.

seed 4 is best on NQ and worst on FiQA, seed 5 is best on Scifact and worst on Scidocs.[11]. Even seed 10, which is the only model that is worst on more than 2 datasets (it is worst on 6) is still best on TREC-Covid.[12]

Our results show that there is no single best retriever, which both supports the motivation of the BEIR benchmark (to give a more well rounded view on retriever performance via a combination of diverse datasets) and shows the need for more analysis into random initialisation and shuffle.

As an addendum, we note that Sellam et al. (2022) did extensive experiments with both random initialisation and data shuffle, and found initialisation to matter more. We did our own experiments to this effect where we trained five MultiContrievers from the same MultiBert initialisation with different data shuffles, from the best, worst, and middle performing seeds. This additional analysis is in Appendix D.

4.3 Q3: Correlation between Extractability & Performance

We tested for correlations across all datasets and common metrics, and present a selection here (Fig 3). Neither NQ (Fig 3a) nor MSMARCO (Fig 3b) correlate with compression metrics. NQ and MSMARCO are the most widely used of the BEIR benchmark datasets, and we hypothesised them to be most likely to correlate. Both are search engine queries (from Google and Bing, respectively) and contain queries that require occupation-type information (*what is cabaret music?*, MSMARCO) and that require gender information (*who is the first foreign born first lady?*, NQ). However, as the dispersed points on the scatterplots show (Figures 3a, 3b), neither piece of information correlates to performance on either dataset. NQ and MSMARCO are representative; we include plots for all datasets in Appendix F.

This result was somewhat surprising; since the contriever training both regularises and increases extractability of gender and occupation, we might expect this to be important for the task. But perhaps it is relevant for *only* the contrastive objective, and not for the retrieval benchmark. Alternatively, it is possible that this information is important, but only up to some threshold that MultiCon-

[11] This best-worst flip exists for seeds 8, 18, and 23 also.

[12] This is to be taken with a grain of salt - that dataset is interesting for generalisation (as these models are trained on only pre-Covid data), but it is only 50 datapoints. We note also that analysis on seed 13 revealed that seed 10 was also unusual, that analysis can also be found in E.

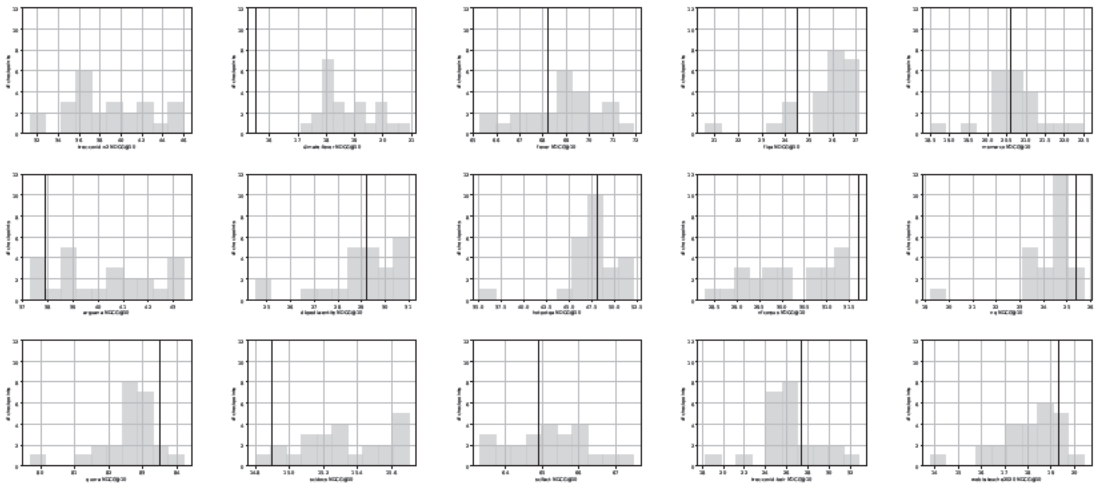


Figure 4: Distribution of performance (NDCG@10) for the 24 MultiContrievers, per BEIR dataset, performance on x-axis, number of models with that value on y-axis. Dashed line indicates reference performance from previous work. While for some datasets the reference performance sits at or near the mean of the MultiContriever distribution, for some the reference performance is a
improvements from random seed can exceed those fi

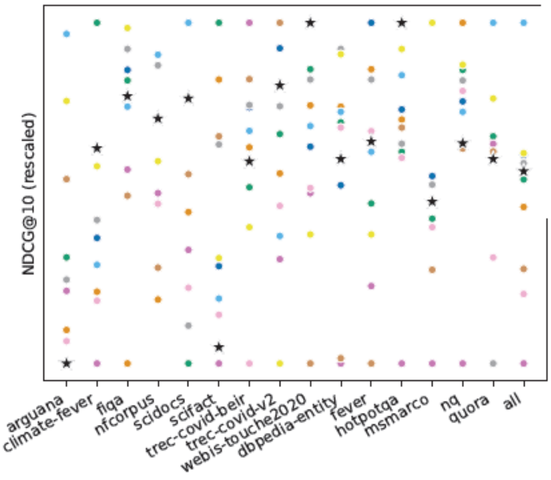
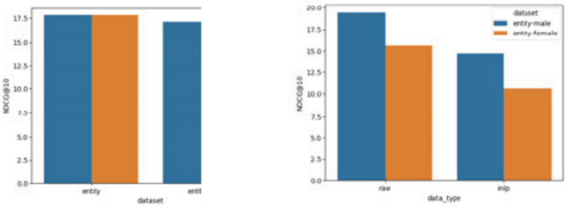


Figure 5: Ranking of best performing seed per dataset (one colour per seed). For legibility, NDCG@10 values are scaled, and all seeds with middle performance are not pictured (10 included). One seed is arbitrarily given a star marker to aid visual interpretation.



(a) NDCG@10 on the neutral vs. gendered NQ entity subsets. Representations are raw (blue) vs. INLP (orange) with gender removed. INLP performance degrades on *only* gender constrained queries: gender is used in those queries, but is not in the control.

(b) Difference in performance between male (blue) and female (orange) entity queries, for raw (left) and INLP (right). The performance gap is constant even when gender is removed via INLP, remains; so the bias is not due to gender in the representations.

Figure 6: INLP experiments

triever models exceed. Finally, it's possible that this information doesn't matter for most queries in these datasets, and so there is some correlation but it is lost, as these datasets are extremely large. This is somewhat supported by the exception cases with correlations being smaller, more curated datasets (F), and so we investigated this as the most tractable to implement.

Our NQ-gender subset of gendered queries (§3.3) does show a strong correlation between gender extractability and performance (Fig 3c). And the NQ-gender subset of neutral non-gendered queries shows no correlation (Fig 3d). So we find that if we isolate to a topical dataset, as e here, extractability *is* predictive of performance, it just isn't over a large diverse dataset.

We strengthen this analysis, testing whether gender information is *necessary*, rather than simply correlated. We use Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020) to remove gender information from MultiContriever representations. INLP learns a projection matrix W onto the nullspace of a gender classifier, which we apply *before* computing relevance scores between corpus and query. So with INLP, the previous Equation 1 becomes:

$$s(d_i, q_i) = \mathbf{W} f_\theta(q_i) \cdot \mathbf{W} f_\theta(d_i) \qquad (3)$$

Then we calculate performance of retrieval with these genderless representations. No drop in performance on the gendered query set with INLP would mean extractable gender information was not necessary. A drop in performance on *both* gender and control queries would support the 'minimum threshold' explanation, or mean that the representation was sufficiently degraded by the removal of gender that other functions were harmed.

Gender information post-INLP drops to 1.4 (nearly none, as 1 is no compression over uniform, Eq 2). Performance on non-gendered entity queries is unaffected, but performance on gendered entity questions drops significantly (5 points) (Fig 6a). From these two experiments we conclude that the increased information extractability *was* useful for answering specific questions that require that information. But most queries in the available benchmarks simply don't require that information to answer them.

4.4 Q4: Gender Extractability and Allocational Gender Bias

Orgad et al. (2022) found gender extractability in representations to be predictive of allocational gender bias for classification tasks; when gender information was reduced or removed, bias also reduced.[13] We found that gender information is *used* (§4.3) so now we ask: is it predictive of gender bias? At least for our dataset, it is not (Fig 6b. This graph shows that there is allocational bias between the female and male queries, and also that the bias remains *after* we remove gender via INLP. *All* performance drops, as we saw for the gendered entities in §4.3. But performance drops by equivalent amounts for female and male entities. These results diverge from what we expected based on the findings of Orgad et al. (2022) for MLMs, who found gender in representations did matter. Our findings suggest that in this case the gender bias comes from the retrieval corpus or the queries, or from a combination. The corpus could have lower quality or less informative articles about female entities (as was found for Wikipedia by Sun and Peng (2021)), or queries about women could be structurally harder in some way.

5 Discussion, Future Work, Conclusion

We trained a suite of 25 **MultiContrievers**, analysed their performance on the BEIR benchmark, probed them for gender and occupation information, and removed gender information from representations to analyse gender bias.

We showed performance to be extremely variable by random seed initialisation, as was the performance ranking of different random initialisations across datasets, despite equal losses during training. Best seed performances often exceed the performance of more complex dense retrievers that use explicit supervision. Future analysis of retriever loss basins to look for differing generalisation strategies could be valuable (Juneja et al., 2023). Our results show that a better understanding of initialisations may be more valuable than developing new models. Our work also highlights the usefulness of metadata enriched datasets for analysis, and we were limited by what was available. Future work could create these datasets and then probe for additional targeted information to learn more about retrievers. This would also enable analysis of demographic biases beyond binary gender.

Gender and occupation extractability was not predictive of performance except in subsets of queries that require gender information. Though both gender and occupation increase in MultiContrievers, the ratio between them decreases, so MultiContrievers should be less likely to shortcut based on gender compared to MultiBerts. We established that the gender bias that we found was not caused by the representations, as it persists when gender is removed. Future work should test in a pipeline is best to correct bias, and how various parts interact. This work also shows the utility of information removal (INLP, others) for causality and interpretability, rather than just debiasing. More availability of test sets for shortcutting could increase the scope of these preliminary results.

Finally, we have analysed only the retriever component of a retrieval system. In an eventual retrieval augmented generation task, the retrieval representation will have to compete with language model priors. The generation will be a composition between unconditionally probable text, and text attested by the retrieved data. Future work could investigate the role of information extractability in the full system, and how this bears on vital questions like hallucination in retrieval augmented generation. We have done the first information theoretic analysis of retrieval systems, and the first causal analysis of the reasons for allocational gender bias in retrievers. We re-

[13] Orgad et al. (2022) use a lexical method to remove gender, but we chose INLP as a more elegant, extensible solution. We replicated their paper with INLP, showing equivalence.

lease our code and resources for the community to expand and continue this line of enquiry. This is particularly important in the current generative NLP landscape, which is increasingly reliant on retrievers and where understanding of models lags so far behind development.

6 Limitations

This work is limited by analysing only one architecture of dense retriever; we chose to experiment instead with random initialisations and shuffles rather than different architectures, so we focused on only the most popular one. So these results may not generalise to all retriever architectures. Our analysis covered only English, and there is work that shows that gender is encoded in a more complex way in other languages (Gonen et al., 2022). INLP, the method we used for causal analysis, is linear, so it might not even work beyond English, though there are recent non-linear extensions of it (Iskander et al., 2023) that could be used in future work.

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *NIPS*.

Maria De-Arteaga, Alexey Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. C. Geyik, K. Kenthapadi, and A. Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Hila Gonen, Shauli Ravfogel, and Yoav Goldberg. 2022. Analyzing gender representation in multilingual models. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 67–77, Dublin, Ireland. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 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)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy J. Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. 2023. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5961–5977, Toronto, Canada. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. 2023. Linear connectivity reveals generalization strategies. In *The Eleventh International Conference on Learning Representations*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Anja Klasnja, Negar Arabzadeh, Mahbod Mehrvarz, and Ebrahim Bagheri. 2022. On the characteristics of ranking-based gender bias measures. In *Proceedings of the 14th ACM Web Science Conference 2022*, WebSci '22, page 245–249, New York, NY, USA. Association for Computing Machinery.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. Predicting inductive biases of pretrained models. In *International Conference on Learning Representations*.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Navid Rekabsaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2065–2068.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. 2021. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander Nicholas D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick, editors. 2022. *The MultiBERTs: BERT Reproductions for Robustness Analysis*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on Wikipedia. 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)*, pages 350–360, Online. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of* *the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg type ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 25–54, Princeton, NJ, USA. PMLR.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

A Contriever Training

Each MultiContriever model was initialised from a MultiBert checkpoint for each of the 25 seeds from 0 - 24, accessed at `https://huggingface.co/google/multiberts-seed_X` where `X` is an integer from 0 - 24. NB: MultiBerts released many checkpoints to enable study of training dynamics, we use only the final complete checkpoint.

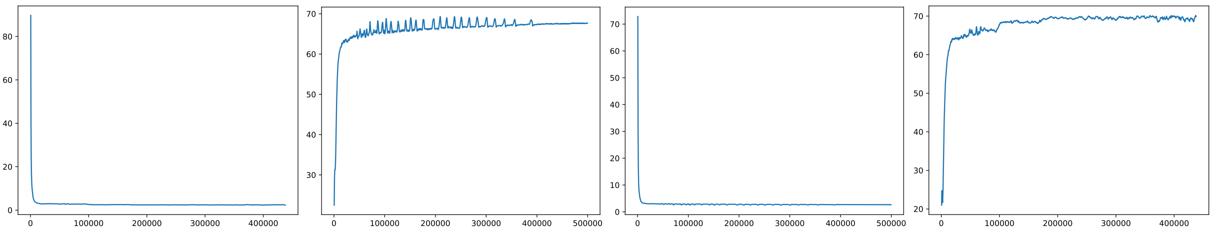
Hyperparameters and training regime is exactly matched to the original Contriever work of (Izacard et al., 2022). Hyperparams can be found in Table 1. Data used was identical to in (Izacard et al., 2022) (from 2019) and was a 50/50 CCNet Wikipedia split.

Each MultiContriever was trained across 4 nodes with 8 GPUs per node (32 GPUs total) for on average 2.5 days. Each MultiContriever was trained for the full 500,000 steps, and checkpointed often; but in all but one seed the best performing checkpoint was the final one (so for that one we use the model at 450,000 steps). This is excepting seed 13, which was anomalous in many other ways (see E).

All MultiContrievers have similar loss and accuracy curves, with seeds 12 and 13 excerpted in Figure 7. All models steeply increase accuracy/decrease loss within 10,000 steps, and then asymptotically approach 69% accuracy by 50,000 steps.

sampling coefficient	0
pooling	average
augmentation	delete
probability_augmentation	0.1
momentum	0.9995
temperature	0.05
queue_size	131072
chunk_length	256
warmup_steps	20000
total_steps	500000
learning_rate	0.00005
scheduler	linear
optimizer	adamw
batch_size (per gpu)	64

Table 1: Hyperparameters used for training MultiContrievers.



(a) Loss (12)

(b) Train Accuracy (12)

(c) Loss (13)

(d) Train Accuracy (13)

Figure 7: Loss and accuracy for seeds 12 and 13, steps on x-axis and loss or accuracy on y-axis.

B Probing Datasets

We rely on two datasets. The first is BiasinBios (De-Arteaga et al., 2019), which is a dataset of web biographies labelled with binary gender, and biography profession. We use De-Arteaga et al. (2019)'s train/dev/test splits of 65:10:25, yielding 255,710 train 39,369 dev, and 98,344 test datapoints. Second is the Wikipedia slice of the `md_gender` dataset (Dinan et al., 2020). This has only labels for gender, which we restrict to be binary since non-binary gender is so small and would adversely affect this analysis. We filter out texts below 10 words (words, not tokens) leaving a dataset of size 10,681,700, split 65:10:25 into 6,943,105 train, 934,649 dev, 2,803,946 test. For practical reasons, we shard it to 9 shards (650,000 train examples each) and then check the results on each shard. All shards behaved consistently. As noted in the text, BiasinBios is nearly balanced with regard to gender labels, but Wikipedia is severely imbalanced.

For both datasets, we use the train set for probing, and the test set for measuring accuracy on the final probe. We investigated using other datasets, but none were of sufficient quality that they were usable. We tested usability very simply: each of the authors labelled a different random sample of 20 examples by hand, and we measured accuracy of dataset labels against our labels, and only took datasets with over 80% accuracy, since our probing task is sensitive to errors in labelling. No other subsets of md_gender nor external datasets that we surveyed passed this bar. We didn't multiply annotate as we found no examples to be at ambiguous.

C Annotation of NQ gender subset

To do our experiments we create a subset of Natural Questions, **NQ-gender**.

We subsample Natural Questions to entity queries by filtering automatically for queries containing any of `who, whose, whom, person, name`. We similarly filter this set into gendered entity queries by using a modified subset of gender terms from Bolukbasi et al. (2016). From this we get a set of queries that is just about entities *Who was the first prime minister of Finland?*, and gendered entities (a female query is *Who was the first female prime minister of Finland?* and a male query is *Who was the first male prime minister of Finland?*).

This automatic process is low precision/high recall. It captures queries with gendered terms in prepositional phrases, (`Who starred in O Brother Where Art Thou?`) which are common false positives in QA datasets, as they are not about brothers. So we manually filter these results by annotating with two criteria: gender of the subject (male, female, or neutral/none (in cases where the gender term was actually in a title or other prepositional phrase as in the example), and a binary tag with whether the query actually *constrains* the gender of the answer. This second annotation is somewhat subtle, but very important. For example, in our dataset there is the query `Who was the actress that played Bee`, which contains a gendered word (actress) but it is not necessary to answer the question; all actors that played Bee are female, and the question could be as easily answered in the form `Who played Bee?`. Whereas in another example query, `Who plays the sister in Home Alone 3?` the query does constrain the gender of the answer. We annotated 816 queries with both of these attributes, of which 51% have a gender constraint, with a gender breakdown of 59% female and 41% male.

We do this annotation ourselves (two of the authors), and we throw out examples that we don't agree on. We are not a representative sample of people (we are all NLP researchers after all) but we consider this lack of diversity to be acceptable since we are not making subjective judgments but are just providing metadata labels.

It is also worth mentioning that two very different types of gender bias in retriever works do create artifacts also, but they are unsuitable for our type of analysis for the following reasons. Rekabsaz and Schedl (2020) and Klasnja et al. (2022) release subsets of MSMARCO, which we did examine and use in initial tests early in this work. Those works define bias very differently, as the genderedness of retrieved documents based on lexical terms, making the implicit normative statement that lack of bias means equal representation of male and female documents in non-gendered queries. This is essentially an independence assertion from fairness literature (Barocas et al., 2019). This is quite different to our approach, which looks at performance disparity between queries that require male and female gender information to answer. Our approach has more immediate practical utility for a real world retriever, and also ties in to the work on information theory by restricting to queries that require gender information. So the lexical document based approach cannot be adapted for our purpose.

D Data Shuffle Experiments

We wanted to answer the question of *If you begin from a worse random initialisation, can you fix it via data shuffle?*. This is of significant practical utility to researchers, who often cannot retrain an existing model from scratch before adapting it to their purpose. Figure 8 shows the best, worst, and a middle performing seed with five additional different data shuffles, and the variance in performance over the datasets. We can see that the worst performing seed is characterised by high variability overall, and the best seed by low variability. So the overall picture is that, on average, the different initialisations determine the quality of the retriever more than the data shuffle. This is in agreement with the findings of Sellam et al. (2022) for MLMs. However, variability is sufficiently high enough that you could get lucky and get the best performance from varying the shuffle, if that is the option available. It would be valuable to extend these to explicit generalisation tasks and interpretabilty challenge sets to see if the high performing shuffles of very variable seeds can be trusted in all settings.

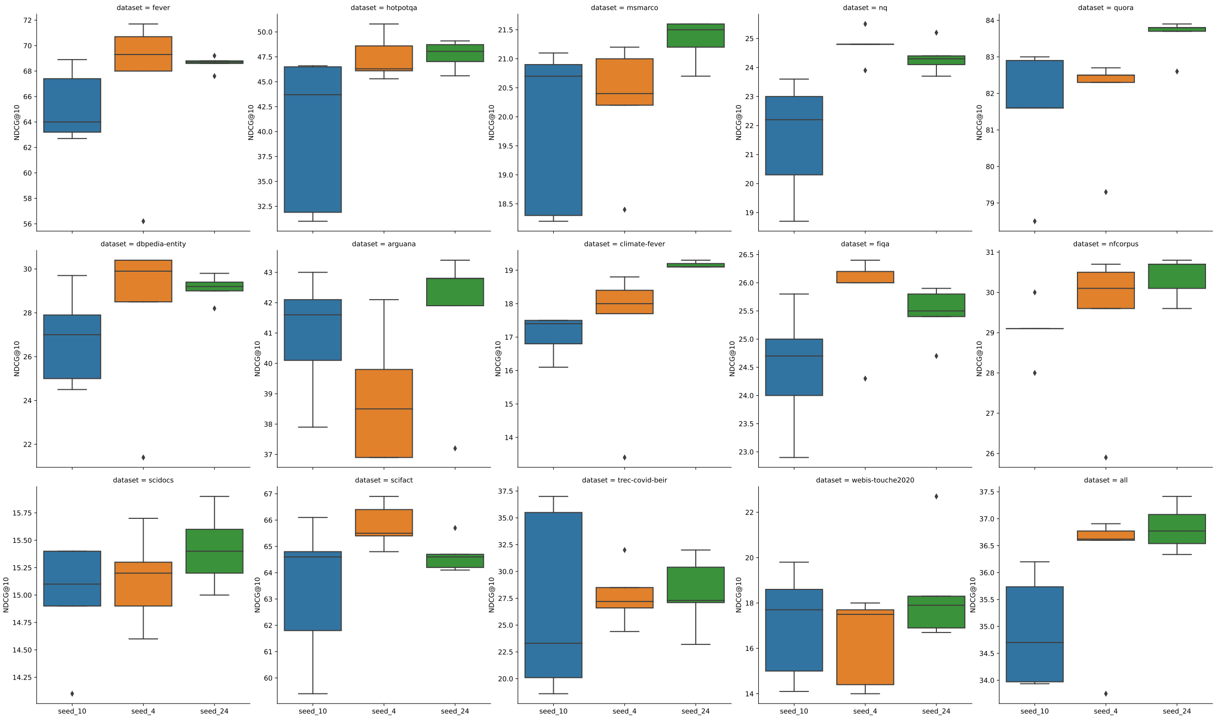


Figure 8: Performance for 5 random datashuffles for a fixed MultiBERT seed - the worst, the best, and a middling seed based on previous experiments. This answers the question of how much variance comes from the random initialisation of parameters, and how much from the data shuffle. It also answers the practical question of 'if you are fine-tuning one model, are you doomed based on the state of the initial model?' The answer is, sort of, but not entirely.

E Seed 13

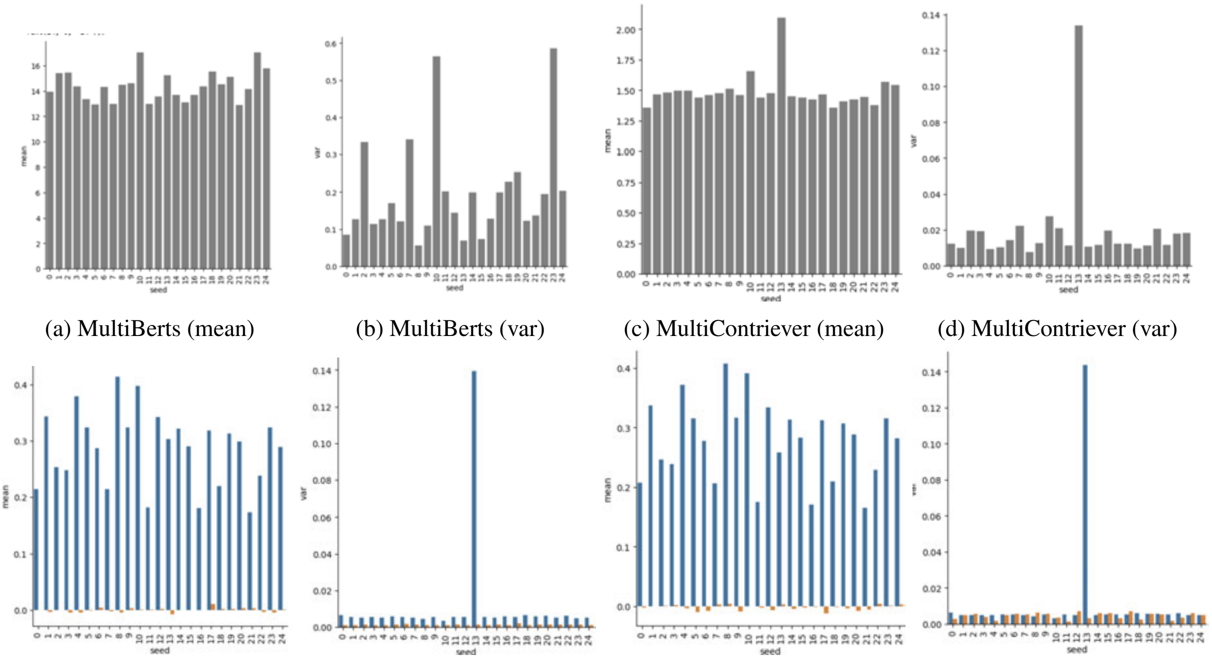
MultiContrievers were trained with seeds 0-24 based on respective MultiBerts 0-24. Seed 13 was excluded from all analysis as it displayed repeatedly anomalous behaviour. During the course of contriever training it appeared indistinguishable from other seeds, loss curves looked normal, there were no signs of overfitting. Performance converged to the same level as other MultiContrievers. However, when applied to the datasets of the BEIR benchmark it did not perform at all, with NDCG of between 2 and 20 on each dataset. We retrained once to replicate the behaviour, and then twice more with different seeds for data shuffle, with identical results. We thus exclude it from all analysis. To aid in future investigations we include our initial analysis of seed 13 irregularities here. We follow the method of analysis of representation spaces from Ethayarajh (2019). We measure the L2 norm of all representations in the BiasinBios dataset (272k) as well as average self-similarity of 1000 randomly sampled representations of those bigraphies, as measured by cosine similarity and by dot product. The former answers the question of how much volume the representations occupy, the latter describes the vector space via how conical (anisotropic) or spherical it is.

In Figure 9, we observe that the vector space of MultiContriever 13 is both larger volume and more obtusely anisotropic (i.e. it occupies a wider cone) than other MultiContrievers. The more obtuse anisotropy originates from MultiBert 13, as can be seen in the high variances for both seeds in cosine similarity. But the larger relative volume happens during the training of the MultiContriever and is unique to it. For MultiBert 13, L2 norm is within normal range, and the anomalous seeds are seeds 10 and 23, which both have larger norms and 5x the variance of other seeds. MultiContriever 13, however, has 1.5x the average norms of all other seeds (which have regularised and become closer in values) and 6x the variance of others. Both MultiBert 13 and MultiContriever 13 have very high variance to average cosine similarity, where the effective range of MultiContriever 13 is -0.03 to 0.53, and MultiBert 13 is 0.02 to 0.58, as compared to other models have a range of 0.28-0.32, for both types of models.

We hypothesise that this reveals a limitation of reliance on the dot product for retrieval, any operation reliant on the dot product loses information when there is a chance of a cosine similarity of zero. We leave other investigation – such as why this would persist from a difference of only random seed initialisation, or why this issue would appear in retrieval, but not in any tasks in the MultiBerts paper, or in the contrastive training process – to future work.

We also note that seed 10 was anomalous in performance compared to the other seeds on the BEIR benchmark; not so anomalous as to be excluded, but it was reliably performing poorly. We can see the higher variance in L2 norms for 10 and 23 in MultiBerts, and then for 10 still in MultiContriever (though nothing noticeable in cosine similarity). Seeds 10 and 13 were not found to be anomalous by Sellam et al. (2022), but they did find seed 23 to display strange behaviour and be extremely unbiased (or even anti-biased) on the Winogender benchmark.

We hope that future work will use our models and continue this line of analysis.



(a) MultiBerts (mean)

(b) MultiBerts (var)

(c) MultiContriever (mean)

(d) MultiContriever (var)

(e) MultiBerts (mean)

(f) MultiBerts (var)

(g) MultiContriever (mean)

(h) MultiContriever (var)

Figure 9: Top row: mean and var of L2 norms of the full BiasinBios dataset for all MultiBert and MultiContriever seeds. Bottom row: mean and var cosine similarity between 1000 random biographies.

F Full set of results for correlation between extractability and performance

Full set of correlations between gender compression and performance in Figure 10 and between profession compression and performance in Figure 11. The latter (profession correlation) have misleading regression lines as only three of 24 models had large differences in compression, such that the line is based off insufficient datapoints. It is included for completeness but left out of analysis for that reason. Gender compression numbers (Figure 10) are distributed more evenly. There are four statistically significant correlations (referred to as by row 1-4, and column a-d, such that the upper left cell is 1a and the lower right cell is 4d). Arguana (1a), Scifact (2b), Webis-Touche (3a), and NQ (4b). All have middling correlation coefficients: Arguana -0.41, Scifact 0.41, Webis-Touche 0.31, NQ 0.42. There is also little in common between these datasets, Arguana and Webis-Touche are argumentation, Scifact is fact-checking, and NQ is google-search style questions. As this leaves most datasets with no correlations, we consider the correlation overall to be weak. We do note that the temporal generalisation datasets are overrepresented in this set (Webis-Touche and Scifact), but leave an investigation of that for future work.

Arguana in particular is unique in having a significant *negative* correlation. We have no answers as to why this might be. It may be a fluke due to peculiarities of this dataset: the dataset is small (less thank 2k datapoints), and is not structured in the same way with query (input) and passage (retrieved) but instead uses a full document passage as the query. It is unclear why this might cause a deterioration in performance from better gender or profession encoding (as we observe the same in profession compression). The Arguana task should match the unsupervised training much more closely since they both are matching the relevance of to document chunks. We leave an investigation into the peculiarities of that dataset also to future work.

G Additional metrics

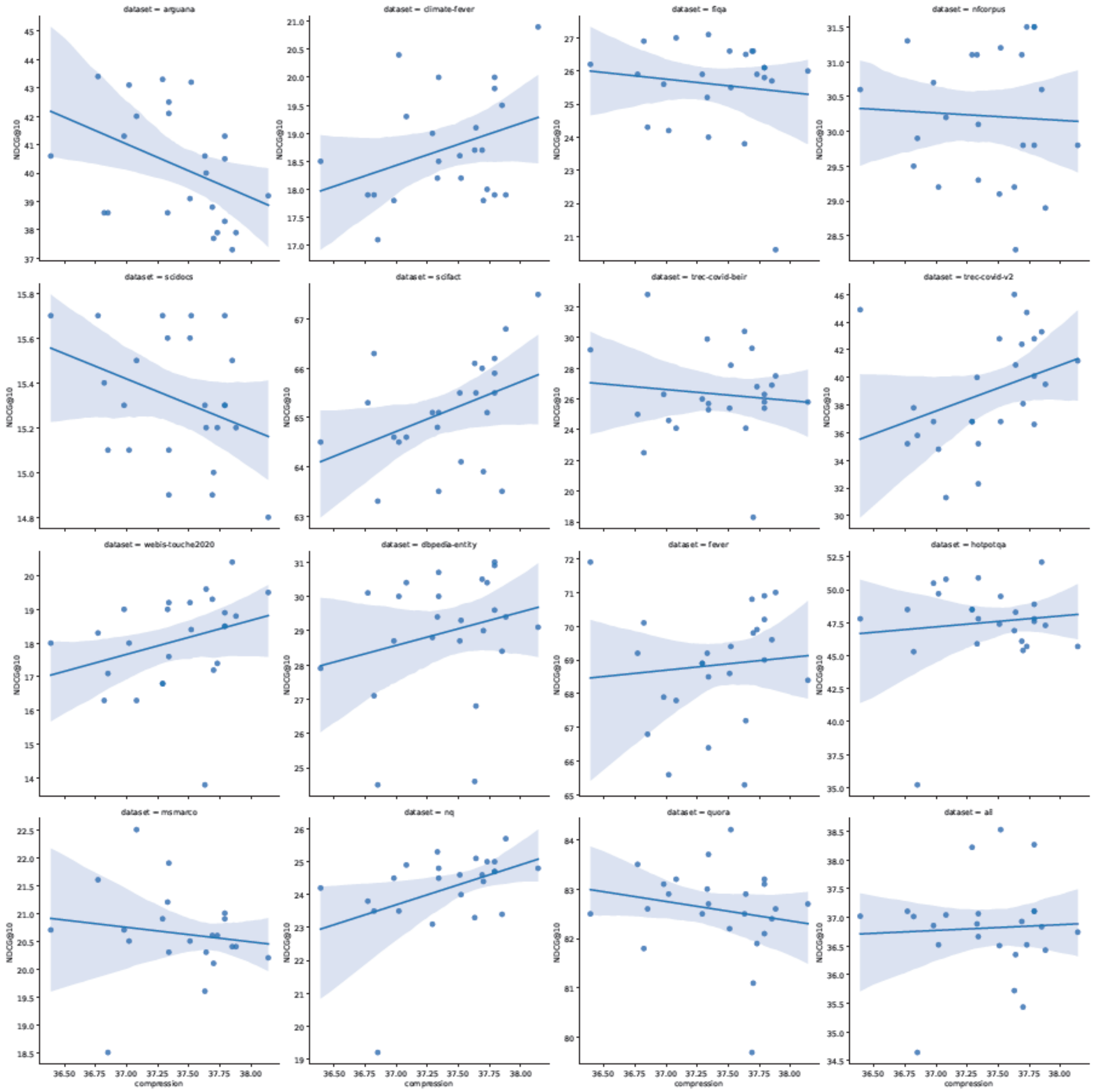


Figure 10: Full set of scatterplots of the correlation between x-axis **gender** compression (ratio of uniform to online codelength) and y-axis performance (NDCG@10), for all datasets individually, and for the average of all BEIR datasets (lower-right). Shaded region is 95% confidence interval.

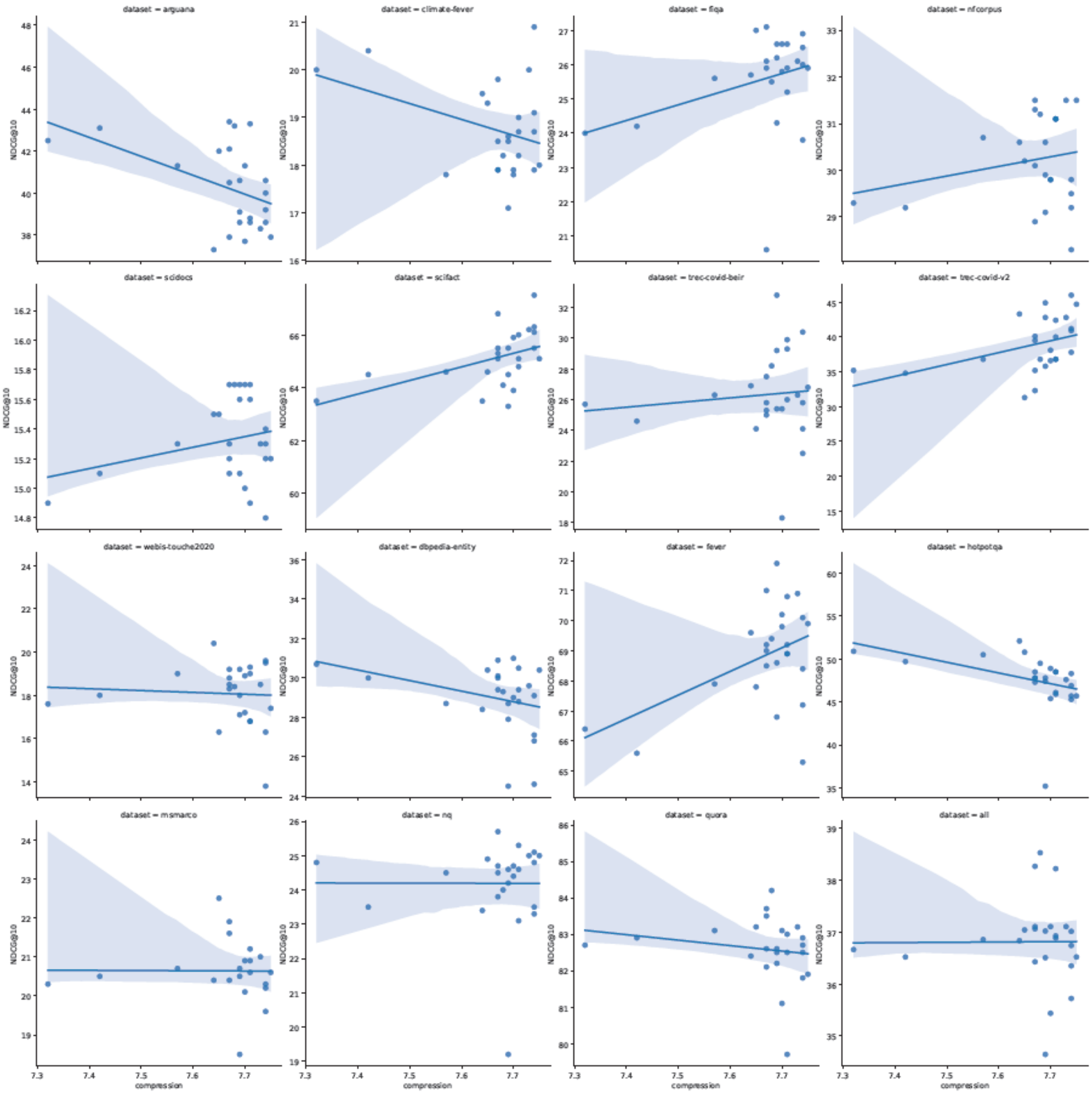


Figure 11: Full set of scatterplots of the correlation between x-axis **profession** compression (ratio of uniform to online codelength) and y-axis performance (NDCG@10), for all datasets individually, and for the average of all BEIR datasets (lower-right). Shaded region is 95% confidence interval.

Part IV

Conclusion

Chapter 8

Conclusions, Questions, and the Present Day

"*For every subtle and complicated question, there is a perfectly simple and straightforward answer, which is wrong.*" —H.L. Mencken

In this section I will condense the takeaways from my work, then expand them into the new questions they pose to the field. I'll discuss implications for future work, do a bit more reflection than in each individual piece of research, made possible by the aggregation of all the my work. I will also include a less traditionally academic section, and relate my work to the present day mania for Large Language models (LLMs), generative AI, and scaling.

8.1 Takeaways

This body of work has focused on **measurement of fairness**, and then, with respect to those measurements, on **analysis of monolingual and cross-lingual transfer**, and finally analysis of **dense retrievers**. It contributes to enriching our understanding of fairness within a multi-part system, which was previously poorly understood. This poverty was much more marked at the commencement of this work than today, but does still persist. Fairness and interpretability research have grown exponentially, but so has the field as a whole, at an equal pace.[1] The scale and speed of productionisation

[1] Fairness work has grown exponentially, but so have ACL and NeurIPS, both with about 40-50% growth in submissions year over year, https://aclweb.org/aclwiki/Conference_acceptance_rates and https://medium.com/criteo-engineering/

still far outstrips our understanding of NLP systems.

In Part I I first examined the dominant method of bias evaluation in language models, based on embedding geometry and cosine similarity, and found it to have poor predictive validity. I advocated for measuring bias downstream. I next found an alternative upstream measure in information theoretic probing of demographics. This measure shows though only *potential* for social bias downstream, a potential that may or may not be realised depending on classifier training. This both enables better understanding, but also reinforces that the full NLP system is still needed to accurately measure fairness.

In Part II, I examined monolingual and cross-lingual transfer in the setting of sentiment classification, in five languages, and found that both settings changed fairness outcomes. Monolingual transfer generally improved fairness despite that this introduces reams of foul pretraining data scraped from the depths of Common Crawl. I attributed this to increased stability in model decisions from the additional data: where stability is defined as not just fewer errors under the counterfactual, but smaller magnitude ones.

Multilingual transfer, however, often worsened fairness outcomes, despite using more data than monolingual transfer. This may not contradict the previous result, as there are less parameters per language, such that even with regard to performance multilingual transfer models are not more stable than monolingual transfer models. The reasons for this difference are left to future work.

In Part III, I examined dense retrieval models, using the tools and research questions accumulated in Parts I and II. I found that random initialisation and random data shuffle play a much larger role than previously thought, and that both performance and fairness were quite sensitive to them. This challenges the standard practice of using only one model with one initialisation and data shuffle for research. Using only one model was then common and is now ubiquitous in the age of LLMs, where training one model takes over 400,000 kWh (Luccioni et al., 2023), or the same amount of energy to make 14 million cups of tea, approximately the same amount that is drunk in Scotland daily.[2]

neurips-2020-comprehensive-analysis-of-authors-organizations-and-countries-a1b55a08132e).

[2] It takes about 336000 joules to raise 1 liter (4 cups of tea) of water to a boil, which is equivalent to 0.116 kWh assuming an electric kettle at 80% efficiency. This makes an LLM equivalent to approximately 13,714,286 cups of tea. The UK drinks on average 3 cups of tea per day (ITC, 2024), and the population of Scotland is 5.4 million today.

In this work I also found a case where information theoretic probing was *not* predictive of gender bias in retrievers, because the bias was caused by other factors beyond the model representation itself.

The combined work in this thesis repeatedly shows how little it makes sense to make choices or come to conclusions about fairness without understanding and simulating the entire system.

8.2 Questions

Each individual work raised as many new questions as it answered. These questions are significant in both size and importance and would be valuable extensions to the understanding of the field, even in the era of LLMs.

For both works in Part I the question remains: what about generative models? What is the relationship between upstream language model metrics and downstream bias in generation, as opposed to in classification? The objective in generation is more similar to that of pre-training, so there is a chance that there is a more predictable correspondence between the two stages. Chapter 4 additionally showed that whether bias was realised was a property of both the language model and the classifier, so what if there's no classifier? Future work could use some of the very recent progress in measuring biases in generation and measuring representational biases (discussed in §2.2) to answer these questions.

In this section (Part I) for classification, I used **observational** studies to determine **allocational bias** by measuring whether the error rates were equal across different demographic populations (male group vs. female group, etc). In Part II I used **interventional studies** that measured whether a change in demographic variable changed predictions, when the predictions should be equal. Both of these measurements require a notion of **equality** — which is very easy with a discrete label space or ordinal values. The same type of study could and should be done for generative models: instead of classifying resumes, a model could write summaries of resumes with a recommendation to proceed or not, which is then read by a human.[3] We can make the same invariance assertion as we made in classification: if we change the gender or race on the resume, the summary should not materially change. But what is a **material**

[3] This is what is happening in practice with generative AI now, as I will discuss in the next section about my experience in Industry.

change? In principle, it is a change that is large enough to cause the summary to be less accurate. Or to be equivalently accurate but to affect a human's opinion positively or negatively. Both would be good operationalisations for different contexts. How do we measure this?

There is scarce previous work on counterfactuals in generative systems, but it raises just as many questions. Vig et al. (2020) consider social bias in generation to be the relative probability of professions like *doctor* and *nurse* under a counterfactual where male and female pronouns are swapped. How would that be extended to different grammatical systems, like Turkish, which has no gendered pronouns? What about different demographic biases that aren't encoded the way gender is? There is far less research about other demographic biases in this kind of setting. The comparatively thin coverage has been noted for race (Field et al., 2021), and there is vanishingly little on other demographic biases (with some exceptions of different coverage, such as Hutchinson et al. (2020), and of very broad coverage, such as Esiobu et al. (2023)). But though they are less well represented in NLP, they are no less important from the viewpoint of ethics or of law.

The second work in Part I, Chapter 4, raises the question: what about beyond gender? Most work in this thesis by design looks at bias beyond gender (3, 5, and 6 all include some notion of race or country of origin) but this one, which proposes a new metric, looks only at gender (partly for lack of suitable datasets beyond it). But gender is encoded very differently in language than other demographic features, so it could reasonably have a different way it operates in model representations and social bias. In English, which weakly marks gender, and other languages with stronger gender agreement, gender information is necessary for correct grammar. A model will need to represent gender well for correct language reconstruction of any text from a noising objective, which is how Transformer models are currently trained (Liu et al., 2019; Lewis et al., 2020a; Vaswani et al., 2017). But race and country of origin are not as strong signals. It is not easy to determine these save from specific words like names,and even then the signal is not as clear as with pronouns, as names do not *just* encode race but also class, gender, time period, etc. How does this difference in encoded information affect the relationship between language models and downstream bias?

In Part II, Chapter 5, we found that there was less bias in aggregate in monolingual transfer, and more reasonable patterns of bias, evidenced by less dramatic changes in sentiment score under the counterfactual. But what about tracing individual examples through from pre-training? Could we track a specific negative stereotype in pre-training and see if it affects decisions later? Extending to the work in Chapter 6, could we extend tracing individual biases into multiple languages?[4] Almost all bias research is done on aggregate information, and we extended our focus to be on patterns of bias, but we stopped short of doing fine-grained analysis, which would be valuable.

We've spent this thesis tracing how fairness persists and travels through a system at a macro level, but we could extend this to a micro level. Such research would not even be bias specific; for there isn't concrete knowledge yet of how *any* information travels between different training stages of models (of which there are increasingly many in the age of LLMs).[5]

In Background Section 2.4, I introduced the notion of fairness as a generalisation error vs. as a learnt dataset artifact—whether an artifact from spurious correlation or a historical bias. Investigation into this difference in causes could help enlighten why racial bias can increase with cross-lingual transfer. Is it really compounding biases (stereotypes) or could it be a generalisation error? Can an investigation into model uncertainty help illuminate which of these cases causes the effects we have observed?

Part III shares the question of 'biases beyond gender' from Chapter 4, as it is also solely gender focused. It also raises questions are general to our understanding of NLP systems as a whole, but have particular importance to fairness. Why is random seed initialisation so important for bias and for generalisation? Why is it possible for a couple of seeds to *just not work at all*, never mind fairness? Some of the anistropy of the representations from earlier training stages seems potentially predictive of later behaviour. Can we understand this well enough to utilise it? If so we could potentially be able to actively encourage model training that is less prone to shortcutting.

8.3 Present Day

Now I want to bring this into current industry practice and zeitgeist. I do this partly because I've spent the better part of the last year working full-time on fairness at an LLM company (Cohere). And partly because, in reviewing my PhD work, I don't

[4] One work has recently come out that also shows 'stereotype leakage' across languages (Cao et al., 2023), which also helps form a foundation for this new question.

[5] There are some methods like this that are starting to do this, like influence functions (Grosse et al., 2023) and some methods that try to do this with Natural Language explanations, which is nonsense and does not work, as Huang et al. (2023) found 'no evidence for causal efficacy' of them.

want to ignore the sea change in NLP research that's taken place over the past year and a hald. I am not someone for whom '*scaling is a way of life*'[6], but it would be disingenuous, in a field intended to improve people's lives, to not speak about how my work relates to current research, current discourse, and current industry practice. This thesis was initially inspired by a sea change that I saw happening six years ago, after all.

This work was all done on models three orders of magnitude smaller than the ones that I deal with in my work today.

This does not matter as much as it might seem. No conclusions in this thesis were model specific. If some architecture arises to replace neural embeddings, LSTMs, and Transformers which bears no genetic link to them, then they *may* no longer hold. But until then, the differences between the models I use at work today and the models in this thesis are: 1) scale 2) a veneer of preference tuning (RLHF, DPO, etc) (Rafailov et al., 2024) 3) instruction tuning (Ouyang et al., 2022) and 4) more training data that is explicitly in the domain of math, logic, and code than we used to include in general NLP models.[7] None of these differences affect my conclusions.

In my first rebuttal to ACL reviewers when the work in Chapter 3 was under review, one of the reviewers asked the common reviewer question 'But have you tried this on `Newest Model Architecture`' (which in this case, was BERT). Adam gave me the advice to turn that into a the question: 'Is there any reason to expect that `Newest Model Architecture` would behave differently? Otherwise, they're just saying it is magic'. To answer this question broadly for LLMs: there is no reason to believe any of the four recent innovations change the things we discovered about fairness in this thesis.

Here are some examples of this being proven.

The replication and extension of my work in Chapter 3 by Cao et al. (2022) did use BERT, and 18 other transformer architectures of varying sizes, and came to the same conclusions. We've seen the same bias amplification affects in LLMs at scale (Bianchi et al., 2023) that we saw in small models in Zhao et al. (2017).

Current research shows that RLHF and the family of preference tuning algorithms pre-

[6] This light shade given by Tatsunori Hashimoto when questioned about it at GenBench at EMNLP 2023.

[7] Though some amount of math and code will be present in CommonCrawl, which does drive the models in this thesis.

dominantly affect *style and structure* of generation, rather than content (Min et al., 2022; Lin et al., 2023). More research shows it can be trivially changed with a few dozen fine-tuning examples (Qi et al., 2023) and that it quickly 'wears off' over conversation turns (Touvron et al., 2023). All information is learnt at earlier stages, predominantly pretraining (Zhou et al., 2023). So from this we conclude that preference tuning will not affect our conclusions.

There is no research I have seen that enables inferences on the effect of instruction tuning or the inclusion of more code and math in data, but there's no reason a priori to think they would change fairness behaviour.

There is one salient change that will matter. Language models are trained to compress and then reconstruct the data they were trained on, and this lossy compression has become less lossy as an effect of scaling. That is: LLMs memorise more individual training samples (Karamolegkou et al., 2023). This could change fairness outcomes, though will it help or will it harm? This depends somewhat on whether the source of the unfairness is a dataset artifact or a generalisation error (§2.4). On the one hand, overall increased memorisation is likely to exacerbate the learning of artifacts. On the other hand, we don't yet understand how scaling affects generalisation, as it is too difficult to test in the current era of closed language models and unknown pretraining and fine-tuning data.

Regardless of this, scaling won't affect the measurements or mechanisms of bias transfer. But these potential interactions of scaling do lend weight to the need for more work on disentangling sources of bias and looking at the effects of increased memorisation from overparameterisation. To date almost all work on memorisation has been from the viewpoint of copyright (Karamolegkou et al., 2023), security and privacy (Smith et al., 2023; Hartmann et al., 2023), or rarely, model quality (where memorisation is at odds with generalisation) (Tänzer et al., 2022). The NLP community should also look at it from the viewpoint of fairness.

When I started this thesis I focused on validating metrics, not because of a dedication to evaluation; I had grand plans for applying my ideas to cross-lingual bias mitigation. But I'd seen unvalidated assumptions in the standard metrics of the field, and it made me unwilling to use those metrics in my own work. I didn't want to stake my PhD research on a metric that I didn't trust, and find out 1.5 years in that my intuition not to trust it had been correct. But now that I work on a deployed product, I spend at least half my time on evaluation. Because good evaluation was *always very hard* and the rise of generative AI has only made it harder. And I can only throw darts at a wall (a perhaps unfair caricature of LLM training) if I know when they've hit something useful, and *that's* the hard part, not the dart throwing.

There is some irony in how Chapter 3, my first fairness work, was the seminal work showing that you cannot do upstream social bias mitigation, and then I took a job where I am supposed to do just that. In practice, I need to try, since education about NLP systems is not yet good enough, and the deployers of language models do not yet have the knowledge and resources to do bias mitigation themselves. So I use the tools and discoveries that I made over the course of this thesis to evaluate my models, and measure wide bounds for what types of bias *could* occur in different reasonable settings, and then make this information public, so that deployers know, can work around it, and maybe do something about it.

But this is still not satisfying enough. I do not think we will ever get to a point in which we rely on one single large pretrained model for thousands of use cases and can predict bias effects downstream for anything but the most common ones. All of this research has progressively taught me that I need to consider the entire NLP system in my measurements for bias: the pretraining, the fine-tuning, the task, the inputs, the corpus that a model can query. The limit case of this it that I need to consider the user interface, the users themselves, the societal power structures within which the NLP system is embedded. And I do think, at some stage, these need to be part of NLP experimental conditions. We cannot consider the harmful effects of QA systems providing false information in absence of how it is displayed in a UI, and how much that UI encourages trust or overreliance (Buçinca et al., 2021). Bias research cannot consider stereotypes in absence of the power structures that make them harmful (Blodgett et al., 2021). No more can most NLP systems be considered without these things, which all together make it increasingly complex to predict all of these things at an upstream stage.

But we can get to a point where we understand better the effect of the choices we've made in the life-cycle of an NLP system. Which ones tend to make things worse, which better, and why. With that, we can better predict potential bias in new systems, and then allocate evaluations and mitigation methods accordingly. But first, we need to understand our systems as a whole.

Bibliography

Artetxe, M., Ruder, S., and Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

Bertrand, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., and Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504, New York, NY, USA. Association for Computing Machinery.

Blasi, D., Anastasopoulos, A., and Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of* *the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. (2021). Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

Borenstein, N., Stanczak, K., Rolskov, T., Klein Käfer, N., da Silva Perez, N., and Augenstein, I. (2023). Measuring intersectional biases in historical documents. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2711–2730, Toronto, Canada. Association for Computational Linguistics.

Bras, R. L., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M. E., Sabharwal, A., and Choi, Y. (2020). Adversarial filters of dataset biases. In *International Conference on Machine Learning*.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. 5(CSCW1).

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Cabello, L., Bugliarello, E., Brandl, S., and Elliott, D. (2023). Evaluating bias and fairness in gender-neutral pretrained vision-and-language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8465–8483, Singapore. Association for Computational Linguistics.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186.

Cao, Y. T., Pruksachatkun, Y., Chang, K.-W., Gupta, R., Kumar, V., Dhamala, J., and Galstyan, A. (2022). On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Cao, Y. T., Sotnikova, A., Zhao, J., Zou, L. X., Rudinger, R., and au2, H. D. I. (2023). Multilingual large language models leak human stereotypes across language boundaries.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Corbett, G. G. (1991). *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Crawford, K. (2017). The trouble with bias. (keynote at neurips).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North Amer-*

ican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., and Gupta, R. (2021). Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM.

Dixon, L., Li, J., Sorensen, J. S., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *AIES '18*.

Dixon, T. L. (2017). *A dangerous distortion of our families: Representations of families, by race, in news and opinion media: A study*. Color of Change.

Drozd, A., Gladkova, A., and Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In Matsumoto, Y. and Prasad, R., editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan. The COLING 2016 Organizing Committee.

Entman, R. M. (1992). Blacks in the news: Television, modern racism and cultural change. *Journalism Quarterly*, 69(2):341–361.

Esiobu, D., Tan, X., Hosseini, S., Ung, M., Zhang, Y., Fernandes, J., Dwivedi-Yu, J., Presani, E., Williams, A., and Smith, E. (2023). ROBBIE: Robust bias evaluation of large generative language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814, Singapore. Association for Computational Linguistics.

Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019). Understanding undesirable word embedding associations. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Fedden, S., Audring, J., and Corbett, G. G. (2018). *Non-canonical gender systems*. Oxford University Press.

Field, A., Blodgett, S. L., Waseem, Z., and Tsvetkov, Y. (2021). A survey of race, racism, and anti-racism in NLP. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131.

Fraser, K. C., Nejadgholi, I., and Kiritchenko, S. (2021). Understanding and countering stereotypes: A computational approach to the stereotype content model. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Goldfarb-Tarrant, S., Ungless, E., Balkir, E., and Blodgett, S. L. (2023). This prompt is measuring <mask>: evaluating bias evaluation in language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada. Association for Computational Linguistics.

Gonen, H., Kementchedjhieva, Y., and Goldberg, Y. (2019). How does grammatical gender affect noun representations in gender-marking languages? In Bansal, M. and Villavicencio, A., editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 463–471, Hong Kong, China. Association for Computational Linguistics.

Gonen, H., Ravfogel, S., and Goldberg, Y. (2022). Analyzing gender representation in multilingual models. In Gella, S., He, H., Majumder, B. P., Can, B., Giunchiglia, E., Cahyawijaya, S., Min, S., Mozes, M., Li, X. L., Augenstein, I., Rogers, A., Cho, K., Grefenstette, E., Rimell, L., and Dyer, C., editors, *Proceedings of the 7th Workshop* *on Representation Learning for NLP*, pages 67–77, Dublin, Ireland. Association for Computational Linguistics.

Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., Steiner, B., Li, D., Durmus, E., Perez, E., Hubinger, E., Lukošiūtė, K., Nguyen, K., Joseph, N., McCandlish, S., Kaplan, J., and Bowman, S. R. (2023). Studying large language model generalization with influence functions.

Guo, W. and Caliskan, A. (2021). Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA. Association for Computing Machinery.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hartmann, V., Suri, A., Bindschaedler, V., Evans, D., Tople, S., and West, R. (2023). Sok: Memorization in general-purpose large language models. *arXiv preprint arXiv:2310.18362*.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR.

Hosking, T., Blunsom, P., and Bartolo, M. (2024). Human feedback is not gold standard. In *The Twelfth International Conference on Learning Representations*.

Hosseini, S., Palangi, H., and Awadallah, A. H. (2023). An empirical study of metrics to measure representational harms in pre-trained language models. In Ovalle, A., Chang, K.-W., Mehrabi, N., Pruksachatkun, Y., Galystan, A., Dhamala, J., Verma, A., Cao, T., Kumar, A., and Gupta, R., editors, *Proceedings of the 3rd Workshop* *on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 121–134, Toronto, Canada. Association for Computational Linguistics.

Hovy, D. and Spruit, S. L. (2016). The social impact of natural language processing. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Huang, J., Geiger, A., D'Oosterlinck, K., Wu, Z., and Potts, C. (2023). Rigorously assessing natural language explanations of neurons. In Belinkov, Y., Hao, S., Jumelet, J., Kim, N., McCarthy, A., and Mohebbi, H., editors, *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 317–331, Singapore. Association for Computational Linguistics.

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., Maini, V., Yogatama, D., and Kohli, P. (2020). Reducing sentiment bias in language models via counterfactual evaluation. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Pimentel, T., Christodoulopoulos, C., Lasri, K., Saphra, N., Sinclair, A., et al. (2023). A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Hutchinson, B. and Mitchell, M. (2019). 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 49–58.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Iskander, S., Radinsky, K., and Belinkov, Y. (2023). Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computa-*

tional Linguistics: ACL 2023, pages 5961–5977, Toronto, Canada. Association for Computational Linguistics.

ITC (2024). International tea committee statistics on uk tea consumption (excerpted here, as itc is paywalled). https://www.teaandcoffee.net/blog/31015/the-resurgence-of-afternoon-tea-experience-in-the-uk/.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Jacobs, A. Z. and Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA. Association for Computing Machinery.

Jia, S., Meng, T., Zhao, J., and Chang, K.-W. (2020). Mitigating gender bias amplification in distribution by posterior regularization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2936–2942, Online. Association for Computational Linguistics.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Karamolegkou, A., Li, J., Zhou, L., and Søgaard, A. (2023). Copyright violations and large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore. Association for Computational Linguistics.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR.

Kelly, J. (2023). How companies are hiring and reportedly firing with ai. *Forbes*.

Kelvin, W. T. B. (1891). *Popular lectures and addresses*, volume 3. Macmillan and Company.

Kennedy, B., Tyson, A., and Saks, E. (2023). Public awareness of artificial intelligence in everyday activities.

Kiritchenko, S. and Mohammad, S. (2018). Examining gender and race bias in two hundred sentiment analysis systems. In Nissim, M., Berant, J., and Lenci, A., editors, *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Korbak, T., Elsahar, H., Kruszewski, G., and Dymetman, M. (2022). On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Lalor, J., Yang, Y., Smith, K., Forsgren, N., and Abbasi, A. (2022). Benchmarking intersectional biases in NLP. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Lauscher, A. and Glavaš, G. (2019). Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In Mihalcea, R., Shutova, E., Ku, L.-W., Evang, K., and Poria, S., editors, *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Lauscher, A., Ravishankar, V., Vulić, I., and Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020a). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020b). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., and Choi, Y. (2023). The unlocking spell on base llms: Rethinking alignment via in-context learning.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lovering, C., Jha, R., Linzen, T., and Pavlick, E. (2021). Predicting inductive biases of pre-trained models. In *International Conference on Learning Representations*.

Luccioni, A. S., Viguier, S., and Ligozat, A.-L. (2023). Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253):1–15.

Ma, W., Chiang, B., Wu, T., Wang, L., and Vosoughi, S. (2023). Intersectional stereotypes in large language models: Dataset and analysis. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8589–8597, Singapore. Association for Computational Linguistics.

McCoy, R. T., Min, J., and Linzen, T. (2020). BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In Alishahi, A., Belinkov, Y., Chrupała, G., Hupkes, D., Pinter, Y., and Sajjad, H., editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

McCurdy, K. and Serbetci, O. (2017). Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *WiNLP*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In Vanderwende, L., Daumé III, H., and Kirchhoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mrkšić, N., Vulić, I., Ó Séaghdha, D., Leviant, I., Reichart, R., Gašić, M., Korhonen, A., and Young, S. (2017). Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Ng, A. (2016). Nuts and bolts of building ai applications using deep learning. *NIPS Keynote Talk*.

Nissim, M., Patti, V., Plank, B., and Durmus, E., editors (2020). *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, Barcelona, Spain (Online). Association for Computational Linguistics.

Novikova, J., Dušek, O., Cercas Curry, A., and Rieser, V. (2017). Why we need new evaluation metrics for NLG. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway books.

Orgad, H. and Belinkov, Y. (2022). Choose your lenses: Flaws in gender bias evaluation. In Hardmeier, C., Basta, C., Costa-jussà, M. R., Stanovsky, G., and Gonen, H., editors, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington. Association for Computational Linguistics.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Parasurama, P. and Sedoc, J. (2022). Gendered language in resumes and its implications for algorithmic bias in hiring. In Hardmeier, C., Basta, C., Costa-jussà, M. R., Stanovsky, G., and Gonen, H., editors, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 74–74, Seattle, Washington. Association for Computational Linguistics.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. (2023). Fine-tuning aligned language models compromises safety, even when users do not intend to!

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *Not Even on Arxiv*.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Robertson, S. E. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18.

Ruder, S., Vulić, I., and Søgaard, A. (2022). Square one bias in NLP: Towards a multidimensional exploration of the research manifold. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354, Dublin, Ireland. Association for Computational Linguistics.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Saphra, N., Fleisig, E., Cho, K., and Lopez, A. (2023). First tragedy, then parse: History repeats itself in the new era of large language models. *ArXiv*, abs/2311.05020.

Schwartz, R. and Stanovsky, G. (2022). On the limitations of dataset balancing: The lost battle against spurious correlations. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Findings of the Association for Computational Linguistics:*

NAACL 2022, pages 2182–2194, Seattle, United States. Association for Computational Linguistics.

Sellam, T., Yadlowsky, S., Tenney, I., Wei, J., Saphra, N., D'Amour, A. N., Linzen, T., Bastings, J., Turc, I. R., Eisenstein, J., Das, D., and Pavlick, E., editors (2022). *The MultiBERTs: BERT Reproductions for Robustness Analysis*.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2021). Societal biases in language generation: Progress and challenges. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Smith, E. M., Hall, M., Kambadur, M., Presani, E., and Williams, A. (2022). "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Smith, V., Shamsabadi, A. S., Ashurst, C., and Weller, A. (2023). Identifying and mitigating privacy risks stemming from language models: A survey.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Subramanian, S., Han, X., Baldwin, T., Cohn, T., and Frermann, L. (2021). Evaluating debiasing techniques for intersectional biases. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empir-*

ical Methods in Natural Language Processing, pages 2492–2498, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sun, J. and Peng, N. (2021). Men are elected, women are married: Events gender bias on Wikipedia. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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)*, pages 350–360, Online. Association for Computational Linguistics.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA. Association for Computing Machinery.

Tänzer, M., Ruder, S., and Rei, M. (2022). Memorisation versus generalisation in pre-trained language models. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7564–7578, Dublin, Ireland. Association for Computational Linguistics.

Tatman, R. and Kasten, C. (2017). Effects of Talker Dialect, Gender and Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Proc. Interspeech 2017*, pages 934–938.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg,

U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Voita, E. and Titov, I. (2020). Information-theoretic probing with minimum description length. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., and Ordonez, V. (2019). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5309–5318.

Winchcomb, T. (2019). Use of ai in online content moderation. Technical report, Ofcom.

Zhao, J. and Chang, K.-W. (2020). LOGAN: Local group bias detection by clustering. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1968–1977, Online. Association for Computational Linguistics.

Zhao, J., Mukherjee, S., Hosseini, S., Chang, K.-W., and Hassan Awadallah, A. (2020). Gender bias in multilingual embeddings and cross-lingual transfer. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W. (2019). Gender bias in contextualized word embeddings. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech-*

nologies, Volume 1 (Long and Short Papers), pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Zheng, S., Song, Y., Leung, T., and Goodfellow, I. J. (2016). Improving the robustness of deep neural networks via stability training. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4480–4488.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., YU, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., and Levy, O. (2023). LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhou, P., Shi, W., Zhao, J., Huang, K.-H., Chen, M., Cotterell, R., and Chang, K.-W. (2019). Examining gender bias in languages with grammatical gender. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.