Seminar paper: Ethics for NLP
Bias detection and debiasing

Marcel Braasch
Technical University of Munich
Supervised by Tobias Eder
marcelbraasch@gmail.com

1 Introduction

In this paper I will mainly discuss the topics addressed during my seminar presentation. The structure very much follows the given talk and is extended by a discussion about ethical implications of bias in natural language processing (NLP) applications. In Section 2 I will shortly recap the meaning of bias and how it enters the NLP game. Section 4 proceeds with the issues in current dataset creation which leads directly to core of the paper. Section 5 moves on to discussing methods of detecting bias. The former two are based on one of the current research papers in the field. In the subsequent Section 6 I proceed similarly and mainly discuss another recent paper in the field. Finally, these are concluded by Section 7, where ethical concerns are raised and reviewed thoroughly.

2 Bias

Bias in its generic form is manifold. Depending on the respective area and the field of application definitions vary. Though, for almost all of them there is a large intersection. For example, Wikipedia frames bias as a *disproportionate weight* in favor or against an idea (Wikipedia contributors, 2021a). The Oxford dictionary defines it as a *strong feeling* in favor or against people or arguments, while not being objective (Oxford Dictionary, 2021). Becoming more specific, looking at topics which at first seem little NLP related, we observe similar definitions. Cognitive bias is described as a systematic pattern of deviation from a norm (Wikipedia contributors, 2021c). Bias in statistics, for example, refers to the expectation of some statistic differing from its true underlying value to be estimated (Wikipedia contributors, 2021b). Independent of which definition we consult, the meaning of bias points in the same direction for most areas, that is, bias being an inherently normative process where one event, *the biased position*, deviates from another, *the norm* (Blodgett et al., 2020).

3 Bias in NLP

Recent approaches of analyzing bias in NLP applications lend themselves to the prior mentioned deviation from standards. In their meta-study Blodgett et al. (2020) classify each motivation into the following rather coarse-grained categories.

Allocational harms Resources or opportunities are distributed unfairly to some groups. This can often can often observed in, e.g., hiring processes. It is a well known issue that men, given the same or comparable qualifications, are frequently favored over women (Isaac et al., 2009). As long as this societal bias is reflected in data provided to systems, they will learn exactly this skew and likely reproduce unwanted behaviors.

Stereotyping Stereotypes, or clichés, are unfair generalizations being made to specific subgroups. For example, current state-of-the-art language generation models like GPT-2 are known for producing stereotypical results such as predicting a missing word as *fight* in *"Two guys in a bar start a _."*, instead of classifying the missing word as *conversation* (Schick et al., 2021; Radford et al., 2019).

Representational bias Systems may perform worse for certain groups, simply due to lack of diverse data. An example would be a grammatical parser which works well for younger people however, regularly produces faulty results for people older than 70 years. Language is in a constant state of change and likely one group expresses certain circumstances differently than another (Hovy and Søgaard, 2015).

Questionable correlations According to some models, certain words describing a specific group may be semantically more similar to specific negative terms. We often observe this type of bias in word embeddings. Gender or racial bias is a well-known problem in these embedding spaces. Researchers have shown before that, for example, *woman* and *homemaker* share the same relation like *man* and *programmer* (Bolukbasi et al., 2016). Similarly, *black* being related to *criminal* like *Caucasian* is related to *police* (Manzini et al., 2019).

Assigning the respective cases of biases to a class is not clear, but rather fuzzy. Many of the examples shown above could be allocated in more than one class. The preceding definitions serve as a reference point for further analysis. Several use cases as well as consequences of these types of biases in NLP applications are further discussed in Section 7.

4 Annotations Practices

Gaining a general view onto bias in NLP we move on to more specific use cases. Training any type of machine learning model typically requires large amounts of data. Thus, generating big training corpora has become one of the driving forces of progress in NLP. In the early 2000s, shared tasks such as SemEval, CoNLL or SenseEval advanced dataset creation and brought forth such (Sabou et al., 2014). While a selected group experts would annotate samples according to a specific task the predominant paradigm has changed in the last years. Still, a few people annotate datasets, however, the common practice today is hiring a group of crowd workers to create these (Trischler et al., 2016; Williams et al., 2017; Rajpurkar et al., 2016). At best, the amount of workers hired is high and they are a fitting representation of society. Unfortunately, as for the shared tasks, current practices suggest room for improvement. Dataset creation is a costly process, often requiring expert knowledge and lots of time for careful evaluation (Williams et al., 2017), being only one of the few, but the major, issue.

Geva et al. (2019) have investigated three common crowd-sourced datasets, that is **1.** OpenBookQA (**OQA**) by Mihaylov et al. (2018) which is a multiple-choice question answering dataset, **2.** MultiNLI (**MNLI**) by Williams et al. (2017), which is a natural language inference dataset, i.e., a task to recognize textual entailment and **3.** CommonSenseQA (**CQA**) by Talmor et al. (2018), which is a multiple-choice question answering dataset. Important descriptive statistics are show in Table 2. Colors are chosen in accordance to the seminar presentation to ensure easier distinguishing between the datasets.

At first, the number of workers seems promising. E.g., for **MNLI** 387 workers annotated over 400.000 sentences. If samples were created uniformly, this would imply high data diversity. In reality, the distribution is rather skewed and follows a power law. In Figure 2 one can observe annotation skewness.

	MNLI	**OQA**	**CQA**
Samples	400.000	5.500	11.000
Workers	387	85	130
Samples / worker	∼1000	∼65	∼85
90% of samples created by	55% of workers	25% of workers	10% of workers

Table 1: Statistics displaying the skewed data generating process in three common NLP datasets.

One can clearly see that very few workers annotate the majority of the datasets. For **MNLI** the distribution is not as drastic as for the other two, but still far from the an ideal distribution. This type of shift will even be observable in the following experiments, suggesting to review the data generation process.

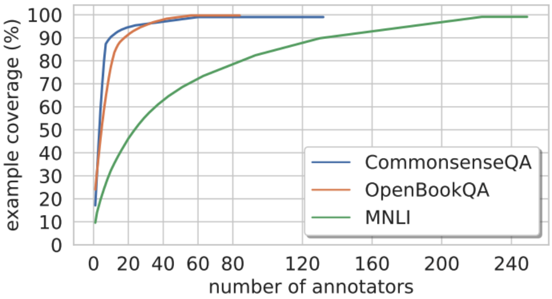

Figure 1: Proportion of examples covered by number of annotators (sorted by number of annotations) after Geva et al. (2019).

5 Detecting Biases

Having typical natural language generation datasets at hand we can proceed to discussing Geva et al. (2019)'s experiments. The experiments are considerably straightforward and quickly lead to conclusions about biases in the underlying datasets. The authors of the paper conducted three experiments. Each experiment tries to prove a hypothesis which may imply bias being present. To be more precise, the hypotheses are: **1.** Given the full annotator information, the model performance increases. **2.**

The model is able to infer the annotator of a given sample. **3.** The model cannot generalize well between annotators. All three experiments strongly indicate that the datasets at hand are biased (in the sense of a representational bias as discussed above). In the following I will briefly showcase the experiments and discuss the results.

1. Utility of annotator information The hypothesis we are testing is simple: given *perfect* annotator information, model performance increases. To verify this, we add the annotator's ID z of the respective creator to every sample (x, y) such that $(x = (z, w_1, \dots, w_{|x|}), y)$. If the model performance increases significantly we can reason that the annotator must induce some latent structure which the model is able to capture. In Table 2 one can observe the absolute gains for each task.

	Without ID	With ID	p-values
OQA	52.2	56.4	$1.83e^{-2}$
CQA	53.6	55.3	$11.98e^{-2}$
MNLI	82.9	84.5	$5.13e^{-7}$

Table 2: Model development performance, after training with/without annotator IDs as additional inputs. p-values were calculated using the McNemar's test (McNemar, 1947) for **MNLI** and the Bootstrap test (Berg-Kirkpatrick et al., 2012) for **OQA** and **CQA**. Results, including this table after Geva et al. (2019).

It becomes apparent quickly that inducing additional annotator knowledge significantly increases model performance. To support their claims, Geva et al. (2019) conducted an additional experiment.

2. Annotator inference For this experiment, the authors test whether the model can infer who created the sample. Instead of doing inference on the original task, the model tries to predict which annotator created a given sample. Formally, samples (x, y), where, e.g., x is a question in **OQA** and y is its respective answer, get replaced with (x, z), where $z \in \{1, \dots, 5, \text{OTHER}\}$. Put simply, we ask the model to infer whether it has been created by one of the top-5 annotators or by the rest. Figure 2 shows the result of the experiment.

We can observe that the model can easily identify the top-5 annotators for **CQA**. Also, it can easily identify the top-1 annotator for **OQA** and struggles to identify its rest. For **MNLI** the model has some ideas of who created the samples but the results are relatively close to the random bottom line.

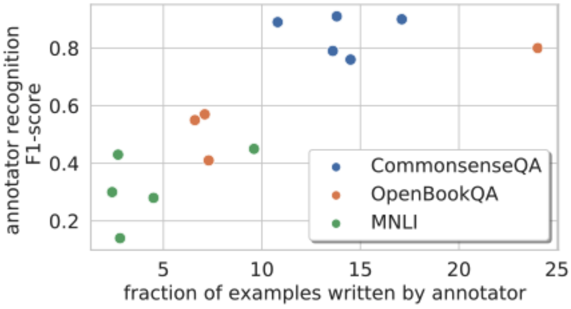

Figure 2: Annotator recognition F1-scores for the top-5 annotators of each dataset (Geva et al., 2019).

3. Generalization across annotators For the last experiment, Geva et al. (2019) constructed the following experimental setting. Draw an example set $\mathcal{S}$ from the entire dataset $\mathcal{D}$. Split $\mathcal{S}$ into two disjoint sets $\mathcal{T} := \mathcal{S} \setminus \mathcal{S}_{\mathcal{Z}}$ and $\mathcal{S}_{\mathcal{Z}}$, where $|\mathcal{S}_{\mathcal{Z}}| \ll |\mathcal{S}|$. The hypothesis now becomes: training the model on the large portion $\mathcal{T}$, the model does not generalize well to $\mathcal{S}_{\mathcal{Z}}$. To become more precise, the authors constructed the following settings.

Multi annotator data split Set $\mathcal{T}$ as the training set where all examples are sampled from the set of examples *without* the top-5 annotators, i.e., random samples from all low-productivity annotators. $\mathcal{S}_{\mathcal{Z}}$ are the samples from the top-5 annotators.

Single annotator data split Set $\mathcal{T}$ the same as before. $\mathcal{S}_{\mathcal{Z}}$ are the samples from only one of the top-5 annotators. For each of the settings run training on $\mathcal{T}$ and inference on $\mathcal{S}_{\mathcal{Z}}$ and observe the results. Figure 3 illustrates the split accordingly.

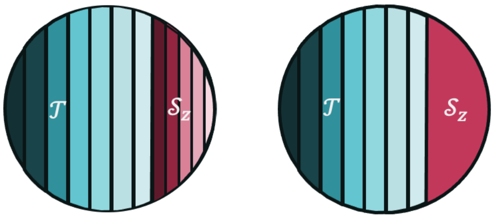

Figure 3: Left: illustration of the multi annotator data split. Right: illustration of the single annotator data split.

As can be seen in Table 3 the results for the multi annotator data split are evident for **OQA**. The model clearly cannot generalize well to the top-5 annotators when being trained on the other, low-productivity, annotators. Recall that **OQA** was the most skewed of all of the datasets, thus the result is not too surprising. For the others, interpretable results are rather uncertain. The performance of

CQA and MNLI dropped in half of the cases, indicating there *might* be a bias present. Similarly, for the single annotator split roughly half of the cases reduced in performance. Note that, for the multi annotator splits test sets where relatively small compared to the training sets (∼1:50), thus experiments where repeated many times to make up for variance and deduce representative results.

Single annotator split

CQA	OQA	MNLI
4.2 ± 0.7	-0.9 ± 2.7	-2.5 ± 0.5
7.7 ± 1.9	-13.5 ± 1.7	-3.0 ± 0.6
-2.8 ± 1.3	-5.8 ± 0.7	-2.9 ± 0.2
-3.8 ± 0.9	8.2 ± 5.2	0.8 ± 0.7
1.6 ± 2.7	3.1 ± 1.1	4.6 ± 0.2

Multi annotator split

CQA	OQA	MNLI
-9.5 ± 8.3	-14.7 ± 6.2	2.5 ± 0.8
6.5 ± 7.0	-19.4 ± 8.5	-1.1 ± 0.9
-6.1 ± 8.5	-12.4 ± 5.5	-4.6 ± 0.8
1.6 ± 10.8	-13.7 ± 8.5	-1.5 ± 0.2
1.8 ± 10.5	-23.3 ± 7.8	0.5 ± 0.2

Table 3: Performance difference between single- and multi-annotator splits. Each cell shows the performance difference mean and its respective standard deviation.

Since results are not very clear, especially for the single annotator split, we may examine the role of each annotator in the data generating process. Geva et al. (2019) hypothesize that specific top-5 annotators create a certain subset of the samples. To be exact, they suspect that the most productive annotators are the same which solve the hard problems. I.e., one could conclude that an issue might be that the test set $\mathcal{S}_{\mathcal{Z}}$ is inherently hard, and possibly *not biased*. To control for inherent hardness the following set up sheds light on the issue.

The underlying hypothesis is that, assuming $\mathcal{S}_{\mathcal{Z}}$ is inherently hard, moving a disjoint portion $\mathcal{P}$ from $\mathcal{S}_{\mathcal{Z}}$ to $\mathcal{T}$ will not increase model performance. If there is some latent structure encoded, i.e., $\mathcal{T}$'s samples are only "hard" because of some annotator bias, then the model performance will increase. The idea is to steadily increase this portion $\mathcal{P}$ and observe model performances.

As can be seen in Figure 4, OQA shows a linear performance increase in F1 score which implies that moving $\mathcal{P}$ over to $\mathcal{T}$ increases model performance and thus the samples are biased, not inherently hard. Similarly, for CQA, at first, model performance increases in a linear fashion and quickly evolves to exponential growth, stating that the bigger we choose $\mathcal{P}$ the quicker the model will learn the induced latent structure encoded by the annotator(s). For MNLI results suggest that samples annotated by the top-5 workers really where just hard samples. Recall that OQA and CQA were the two datasets following the power law. Results are not very surprising for OQA and CQA , thus hypotheses for these are not rejected. For MNLI we likeley reject the hypothesis.

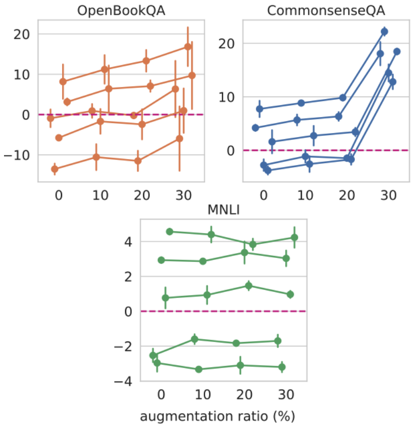

Figure 4: Left: illustration of the multi annotator data split. Right: illustration of the single annotator data split.

To wrap it up, Geva et al. (2019) nicely showcased a method to verify whether a dataset is representationally biased or not. Also, the authors showed that if data variety is not too high, machine learning models can infer who created a sample, increase its performance if this information is provided and often has issues generalizing between annotators. The work presented possibly purports that the dataset creators were not careful enough. However, taking a closer look, one can find that the dataset creation processes at hand follow a carefully selected multi stage pipeline. The creators try to spot outliers and remove discrepancies beforehand. The examined papers are of high quality, but of course, as with most findings, are not flawless.

Clearly, an increase or drop of model performance, depending has a direct impact on real-world applications. Though not as ubiquitous yet, NLP applications are gradually being put to use in daily practice. If systems are being deployed for the purpose of simplifying lives, be it employees of companies or their customers, they must benefit everyone. Just as humans in an human-to-human interaction can adapt, e.g., to a person having a specific accent or using a specific sentence pattern, machines must do so as well. Otherwise, extensive consequences for provider and user are predetermined. For an extending discussion, see Section 7.

Of course, there are other promising approaches to detecting bias apart from the presented one. However, for this seminar, I chose to focus on a specific paper and discuss its method in more depth.

6 Debiasing

Apart from the crowd working paradigm briefly discussed in Section 4 there is a second popular method for retrieving large amounts of data. This type of data is inherently different though, and so is its retrieval process. Here, the data we are interested in is crowd sourced from millions of web pages in order to obtain a huge scale corpus of running text, often comprised of trillions of tokens. Especially with the rise of pre-trained language models, there is an ever-growing need for this type of data (Devlin et al., 2018; Brown et al., 2020; Yang et al., 2019). We do not need annotated data to solve a task but loosely speaking want the model to familiarize itself with the language at hand.

Obviously, this comes with other issues than the ones discussed before. Internet sources such as Twitter, Reddit or Facebook are full of prejudices, hate speech or other harmful formulations (Olteanu et al., 2018; Qian et al., 2019). These types of (often stereotypical) biases can be easily picked up by these powerful models, even if they only occur very few times.

Schick et al. (2021), among others, find that, large language models often reproduce undesirable output. To prevent this behaviour they ask whether these models at least *know* if they have produced a harmful output and call this process *self-diagnosis*. Next, they move on to an approach to reduce the probability to produce undesirable output based on the prior idea, called *self-debiasing*. The authors propose an idea following a few-shot learning approach (Brown et al., 2020). That is, put simple, besides the actual training data we augment the inference (or generation) process by a hint as to what the task is. Pre-trained language models have a "sense of the language" at hand. Current theoretical knowledge about few-shot learning settings augmented by task descriptions is rather vague. But, loosely speaking, if objects of interest, in our case a family of related words embedded in latent space, have a relatively distinct representation, models get a cue in what sub spaces to navigate beforehand (Cao et al., 2019; Tyukin et al., 2021).

Self-diagnosis The idea of self-diagnosis follows the few-shot learning idea. We let the model produce a sentence $\mathbf{x}$ and simply prompt it the template "$\mathbf{x}$; Does the prior text contain $\mathbf{y}$?; Answer __", where $\mathbf{y}$ is a textual description out of the ones which can be seen in Table 4. This template is below referred to as $\text{sdg}(\mathbf{x}, \mathbf{y})$.

Attribute Name	Description
toxicity	rude, discrespectful or unreasonable language
severe toxicity	very hateful, aggressive disrespectful language
sexually explicit	sexually explicit language
threat	a threat
profanity	swear or curse words, or other obscene or profane language
idenity attack	negative or hateful language targeting someone because of their identity

Table 4: Description of the negative attributes which the language model is asked if it produced it.

Next we estimate the probability of $\mathbf{x}$ exhibiting attribute $\mathbf{y}$ as

$$\mathbb{P}(\mathbf{y}|\mathbf{x}) = \frac{\mathbb{P}_M(\text{Yes}|\text{sdg}(\mathbf{x}, \mathbf{y})}{\sum_{w \in \{\text{Yes,No}\}} \mathbb{P}_M(w|\text{sdg}(\mathbf{x}, \mathbf{y}))} \quad (1)$$

where $\mathbb{P}_M$ is the language model at hand and we conclude to have a biased answer if some treshold τ is met, i.e., if $\mathbb{P}(\mathbf{y}|\mathbf{x}) \geq \tau$. The experiments show that with increasing capacity, pre-trained models become well-aware of the harmful outputs they produce (see Figure 5).

Self-debiasing This gives rise to using this internal knowledge to have the model produce harmful outputs with a lesser likelihood. The idea is fairly straightforward. We employ another template $\text{sdb}(\mathbf{x}, \mathbf{y})$, more precisely being "The following text contains $\mathbf{y}$: $\mathbf{x}$ __", given one of the categories shown in Table 4. Using this template, we encourage the model to produce harmful outputs by introducing

$$\mathbb{P}_M(w|\text{sdb}(\mathbf{x}, \mathbf{y})), \quad (2)$$

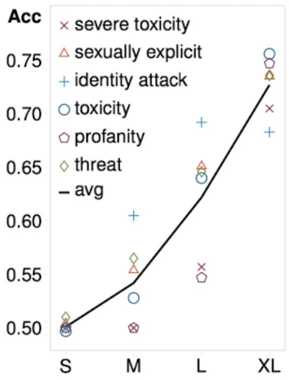

Figure 5: self-diagnosis abilities for the attributes covered in Table 4. GPT-2's sizes are evaluated using accuracy as a metric.

compute the probability of producing the next word by

$$\mathbb{P}_M(w|\mathbf{x}) \tag{3}$$

and continue to define the difference by

$$\begin{aligned}\Delta(w,\mathbf{x},\mathbf{y}) = \quad & \mathbb{P}_M(w|\mathbf{x}) \\ & - \mathbb{P}_M(w|\text{sdb}(\mathbf{x},\mathbf{y}))\end{aligned} \tag{4}$$

Equation (4) will be less than 0 for undesirable terms if and only if the language model's probability (2) is lesser than the toxic language model's probability (3). Using this fact one can introduce a new probability distribution for a debiased language model which mitigates bad outputs as

$$\mathbb{P}_D(w|\mathbf{x}) \propto \alpha(\Delta(w,\mathbf{x},\mathbf{y})) \cdot \mathbb{P}_M(w|\mathbf{x}) \tag{5}$$

where $\alpha : \mathbb{R} \to [0,1]$ is a scaling function to control the influence of the the delta term (for more details on α, please refer to the paper). The authors were able to show that their approach is able to halve toxicity, sexually explicit outputs and profanity. Furthermore, severe toxicity, threats and identity attacks could even be reduced by approximately 60%.

Again, this approach is only one of few promising ideas. Other interesting approaches include Bolukbasi et al. (2016)'s work, which shows that word vectors can be debiased by projection words onto the gender direction by $w \cdot (\vec{she} - \vec{he})$ and adjusting differences between these. Gonen and Goldberg (2019) later show that the former approach rather covers the issue instead of resolving it.

In a rather theoretical manner, He et al. (2019) show an interesting approach by utilizing residual fitting. They fit a model f^s_θ, which will be biased, then fix $f^s_{\theta*}$. In an iterative fashion, they fit same model on f^d_ϕ, which acts like the residual gap of f. I.e., the loss $min_\phi \mathbb{E}[L(f^s_{\theta*} + f^d_\phi, y)]$ is minimized until the gap does not close further.

Bias in NLP, from a philosophical, or rather motivational view, up to really technical, mathematical approaches, is a very lively field in computational linguistics. As we will see in the proceeding discussing, there are many reasons for this.

7 Ethical Implications

Now that we have generally discussed current methods in bias detection and removal we can move on to a general discussion of ethical implications of these. If no citation is visible, the statement reflects my own opinion or knowledge about the field. Besides, my own ideas are augmented by the discussion and ideas brought up during the seminar meeting.

To start off, let us have a look at the dataset generation process. We need to distinguish between datasets annotated by crowdworkers and those sourced from the internet. Looking at the former, one might wonder if there is a need to standardize this process to obtain some sort of quality label. Information such as *who* created the data, with what *purpose* in mind, is there *bias known* or is it possible to *abuse the data* are only a few questions one could pose. Gebru et al. (2018) provide such a framework by introducing questions concerning the dataset creators' motivation, the composition, the collection process and more. This is an important step in the right direction. Researchers do not have to worry about what type of meta-questions to ask about their work. They can proceed to answering the ones, frameworks like these provide. However, there comes a downside with it. As long as these guidelines are not a *must*, they may appear cumbersome. Putting in extra work while not, at least individually, benefiting from it may inhibit researchers from employing this type of framework. Conference deadlines are strict and research groups are already under pressure to meet those. Nevertheless, standardizing and controlling data creation is inevitable if we want controlled and regulated AI. Just like we look for transparent and explainable machine learning models, we need equal measures for their fuel. Fortunately, controlling these processes seems like a practice in demand. Leading conferences such as the Conference on Neural Information Processing Systems (NeurIPS) explicitly ask for an ethical statement concerning the research to be published (NeurIPS, 2021). The Annual Meetings of the Association for Computational Linguistics (ACL) *recommends* to discuss possible ethical concerns found by the researchers. For new datasets the aforementioned framework (among others) is referenced (ACL, 2021). Researcher retain some autonomy without being forced into strict guidelines. This enables utilizing an ethical discussion to one's specific need, however, it may also result in imprecise or even incomparable evaluations.

The former is a rather qualitative approach of setting a dataset creation standard. Besides that, one could look for a general quantitative tool for analyzing the adequacy of existing datasets. By some threshold indicating "biasy", it may be decidable if a dataset can be used to train a model. Clearly, before asking *how biased is this dataset* we need to define what type of bias we are examining. Mehrabi et al. (2021) specify 23 different, though partially fuzzy, types of biases. Further they define 10 different concepts ranging from various types of fairness to concepts such as equal treatment. These definitions are crucial to proceed with quantitative approaches of measuring the extent of biases in datasets. The authors point to various promising sources of further investigation. Once this foundation is carefully laid out one could look out for a general quantitative process, though, I sense that there are more important questions to ask before moving on. A general framework seems like a very hard task, if solvable at all. Unfortunately, a detailed analysis of quantitative bias estimation is out of scope for this seminar work. For the further discussion, I will orient myself towards the coarse-grained concepts defined in Section 3, but wanted to mention that other authors may have laid noteworthy work with this regard.

Especially in the context of word embeddings, bias with respect to *questionable correlations* has been investigated thoroughly in the last years (Sweeney and Najafian, 2019; Chaloner and Maldonado, 2019). The prevailing measure for quantifying demographical bias in word embeddings is WEAT, a statistical measure used for hypothesis testing (Caliskan et al., 2017). This tool has long been in use to counteract tendencies encoded in our language (Gonen and Goldberg, 2019). Questionable correlations can have far-reaching consequences, for those employing biased systems as well as those affected by them. In many countries like Germany, discrimination against individuals due to their age, sex, gender or religious affiliation is illegal and has been penalized many times (Hamm, 2015; Aachen, 2012). Court decisions to date have mostly been consequences of improper behavior conducted by human individuals. However, there are cases where entire systems needed to be shut down due to harmful, and clearly biased, behaviour. This has gone as far as women being systematically discriminated in hiring processes (Dastin, 2018). Word embedding approaches tend to pick up biased correlations such as *women* being related to *homemaker* as *men* are to *programmer*. In some ways, these embeddings reflect societal circumstances present at the moment. However, these trends are conditioned on historical biases in the data (Mehrabi et al., 2021), which similarly, is conditioned on traditional role models as well as a breeze of patriarchism (personal opinion). There are, however, strong endeavours towards breaking old patterns and empowering equality between genders. Many firms endorse diversity and openly encourage prospective employees from the entire population stratum to apply for jobs (Google, 2021; Amazon, 2021). Besides economic damages, which for many are insignificantly small, companies must anticipate reputational risks if they cannot cope with their communicated values, whether through individual human, or automated non-human action. Besides, individuals can take emotional damage through biased or unfair decisions made by an automated system.

Issues of *representational bias*, especially with respect to systems performing worse for a certain subgroup or poor generalizing abilities in general, have been showcased in Section 5. Geva et al. (2019)'s and Schick et al. (2021)'s approaches show fundamental works demonstrating promising ideas of how detection of bias might work but it only scratches the surface of the issue. With these, we can detect *if* a dataset has representational issues and by *what margin*. However, the later is neither comparable nor generalizable between datasets and tasks, if not further standardized. Furthermore, we cannot yet determine *who* exactly is affected. Determining the affected groups yields other inherent problems. Defining certain sub-populations as well as their granularity is troublesome and intrinsically difficult. It requires careful examination augmented by anthropological and sociological domain knowledge.

As can be seen in Section 6 there are approaches for mitigating bias during inference. These are measures taken once the model already has acquired its knowledge. One may wonder whether it makes sense to restrain the model to learn harmful terminology in the first place. If these terms in the crowdsourced datasets are removed beforehand, models cannot pick them up. A fairly naive approach would be removing sentences based on a list of banned words. This would likely remove a large majority of target sentences. While this seems like a reasonable thing to do, it raises several issues. If we restrict models to sentences we want them to *generate*, we restrict them to *understand* these at the same time. In applications like, for example, a BERT part-of-speech tagger this presumably works without limiting performance. However, variants of BERT like these are not relevant for our case. Clearly, we are interested in critical applications with high human-computer-interaction like a conversational or question answering system. Besides, the prevailing paradigm in *multi purpose* pretrained language models is exactly that. We want to use the same model for many tasks and want it to adapt to various situations, just like humans. If we look at general purpose language models such as BERT, we clearly want the model to make sense of its environment. For a conversational system a lack of understanding can entail completely faulty responses. It could even miss to understand dangerous situations where a consequent action is needed, e.g., an antisemitic comment or a (death) threat. In many cases we want to take immediate measures against these. Just by a few examples, it becomes apparent that restricting models of parts of the real world does not seem like an adequate solution. Loosely speaking, it is like trying to protect a child from the actual, sometimes brutal world. We are in need of models as knowledgeable as adults. These models, however, should act accordingly at the same time, that is producing proper language.

To conclude the discussion, I personally believe that language models will keep finding use in our everyday. It is an important development that will ease many of the everyday challenges. But every groundbreaking invention, evidently, comes with a risk. In line with this, we need to make sure that these models are deployed based on (in our case, European) values such as inclusion, tolerance, justice, solidarity and non-discrimination. Many of the points raised before pose a possible infringement of these and it is essential to develop safe technology that not only benefits few.

References

Arbeitsgericht Aachen. 2012. Urteil Diskriminierung: Fehlende Religionszugehörigkeit. bit.ly/3yDAO7t. Online; accessed 08. August 2021.

ACL. 2021. ACL Ethics FAQ. bit.ly/3fVu8ub. Online; accessed 08. August 2021.

Amazon. 2021. Gestaltung einer inklusiven Kultur. bit.ly/37y4UgT. Online; accessed 09. August 2021.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Tianshi Cao, Marc Law, and Sanja Fidler. 2019. A theoretical analysis of the number of shots in few-shot learning. *arXiv preprint arXiv:1909.11722*.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32.

Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Online; accessed 08. August 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.

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. *arXiv preprint arXiv:1903.03862*.

Google. 2021. Build for Everyone. diversity.google. Online; accessed 08. August 2021.

Landesarbeitsgericht Hamm. 2015. Urteil Bewerbungsabsage: Diskriminierung wegen des Geschlechts. bit.ly/3jBW9bg. Online; accessed 08. August 2021.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. *arXiv preprint arXiv:1908.10763*.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, pages 483–488.

Carol Isaac, Barbara Lee, and Molly Carnes. 2009. Interventions that affect gender bias in hiring: A systematic review. *Academic medicine: journal of the Association of American Medical Colleges*, 84(10):1440.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

NeurIPS. 2021. Ethical Guidelines. bit.ly/3yFGrCa. Online; accessed 08. August 2021.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Oxford Dictionary. 2021. Bias. Online; accessed 05. August 2021.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866. Citeseer.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *arXiv preprint arXiv:2103.00453*.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Ivan Y Tyukin, Alexander N Gorban, Muhammad H Alkhudaydi, and Qinghua Zhou. 2021. Demystification of few-shot and one-shot learning. *arXiv preprint arXiv:2104.12174*.

Wikipedia contributors. 2021a. Bias. Online; accessed 05. August 2021.

Wikipedia contributors. 2021b. Bias (statistics). Online; accessed 05. August 2021.

Wikipedia contributors. 2021c. Cognitive bias. Online; accessed 05. August 2021.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.