Predictive Biases in Natural Language Processing Models:
A Conceptual Framework and Overview

Deven Shah
Stony Brook University
`dsshah@cs.stonybrook.edu`

H. Andrew Schwartz
Stony Brook University
`has@cs.stonybrook.edu`

Dirk Hovy
Bocconi University
`dirk.hovy@unibocconi.it`

Abstract

An increasing number of natural language processing papers address the effect of bias on predictions, introducing mitigation techniques at different parts of the standard NLP pipeline (data and models). However, these works have been conducted individually, without a unifying framework to organize efforts within the field. This situation leads to repetitive approaches, and focuses overly on bias *symptoms/effects*, rather than on their *origins*, which could limit the development of effective countermeasures. In this paper, we propose a unifying *predictive bias framework for NLP*. We summarize the NLP literature and suggest general mathematical definitions of predictive bias. We differentiate two consequences of bias: *outcome disparities* and *error disparities*, as well as four potential origins of biases: *label bias*, *selection bias*, *model over-amplification*, and *semantic bias*. Our framework serves as an overview of predictive bias in NLP, integrating existing work into a single structure, and providing a conceptual baseline for improved frameworks.

1 Introduction

Predictive models in NLP are sensitive to a variety of (often unintended) biases throughout the development process. As a result, fitted models do not generalize well, incurring performance and reliability losses on unseen data. They also have socially undesirable effects by systematically underserving or mispredicting certain user groups.

The general phenomenon of biased predictive models in NLP is not recent. The community has long worked on the domain adaptation problem (Jiang and Zhai, 2007; Daume III, 2007): models fit on newswire data do not perform well on social media and other text types. This problem arises from the tendency of statistical models to pick up on non-generalizable signals during the training process. In the case of domains, these non-generalizations are words, phrases, or senses that occur in one text type, but not another.

However, this kind of variation is not just restricted to text domains: it is a fundamental property of *human-generated* language: we talk differently than our parents or people from a different part of our country, etc. (Pennebaker and Stone, 2003; Eisenstein et al., 2010; Kern et al., 2016). In other words, language reflects the diverse demographics, backgrounds, and personalities of the people who use it. While these differences are often subtle, they are distinct and cumulative (Trudgill, 2000; Kern et al., 2016; Pennebaker, 2011). Similar to text domains, this variation can lead models to pick up on patterns that do not generalize to other author-demographics, or to rely on undesirable word-demographic relationships.

Bias may be an inherent property of any NLP system (and broadly any statistical model), but this is not per se negative. In essence, biases are priors that inform our decisions (a dialogue system designed for elders might work differently than one for teenagers). Still, undetected and unaddressed, biases can lead to negative consequences: There are aggregate effects for demographic groups, which combine to produce *predictive bias*. I.e., the label distribution of a predictive model reflects a human attribute in a way that diverges from a theoretically defined "ideal distribution." For example, a Part Of Speech (POS) tagger reflecting how an older generation uses words (Hovy and Søgaard, 2015) diverges from the population as a whole.

A variety of papers have begun to address countermeasures for predictive biases (Li et al., 2018; Elazar and Goldberg, 2018; Coavoux et al., 2018).[1] Each identifies a specific bias and counter-

---

[1] An even more extensive body of work on *fairness* exists as part of the FAT* conferences, which goes beyond the scope
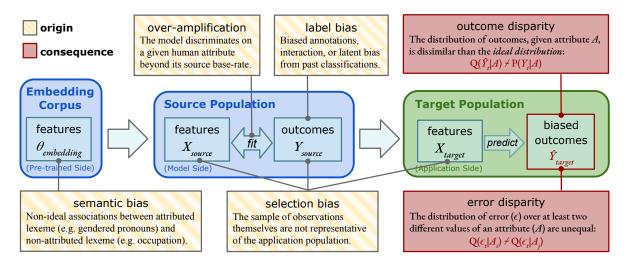
Figure 1: *The Predictive Bias Framework for NLP*: Depiction of where bias may originate within a standard supervised NLP pipeline. Evidence of bias is seen in $\hat{y}$ via *outcome disparity* and *error disparity*.

measure on their terms, **but** it is often not explicitly clear which bias is addressed, where it originates, or how it generalizes. There are multiple sources from which bias can arise within the predictive pipeline, and methods proposed for one specific bias often do not apply to another. As a consequence, much work has focused on bias *effects and symptoms* rather than their *origins*. While it is essential to address the effects of bias, it can leave the fundamental origin unchanged (Gonen and Goldberg, 2019), requiring researchers to rediscover the issue over and over. The "bias" discussed in one paper may, therefore, be quite different than that in another.[2]

A shared definition and framework of predictive bias can unify these efforts, provide a common terminology, help to identify underlying causes, and allow coordination of countermeasures (Sun et al., 2019). However, such a general framework had yet to be proposed within the NLP community.

To address these problems, we suggest a joint conceptual framework, depicted in Figure 1, outlining and relating the different origins of bias. We base our framework on an extensive survey of the relevant NLP literature, informed by se-

lected works in social science and adjacent fields. We identify four distinct sources of bias: **selection bias**, **label bias**, **model overamplification**, and **semantic bias**. We can express all of these as differences between (a) a "true" or intended distribution (e.g., over users, labels, or outcomes), and (b) the distribution used or produced by the model. These cases arise at specific points within a typical predictive pipeline: embeddings, source data, labels (human annotators), models, and target data. We provide quantitative definitions of predictive bias in this framework intended to make it easier to: (a) identify biases (because they can be classified), (b) develop countermeasures (because the underlying problem is known), and (c) compare biases and countermeasures across papers. We hope this paper will help researchers spot, compare, and address bias in all its various forms.

Contributions Our primary contributions include: (1) a conceptual framework for identifying and quantifying predictive bias and its origins within a standard NLP pipeline, (2) a survey of biases identified in NLP models, and (3) a survey of methods for countering bias in NLP organized within our conceptual framework.

2 Definition - Two Types of Disparities

Our definition of *predictive bias* in NLP builds on its definition within the literature on standardized testing (i.e., SAT, GRE, etc.) Specifically, Swinton (1981) states:

> By "predictive bias," we refer to a situation in
> which a [predictive model] is used to predict a

---

of this *biased-focused* paper. Note also that while bias is an ethical issue and contributes to many papers in the ethics in NLP area, the two should not be conflated: ethics covers more than bias.

[2] Quantitative social science offers a background for bias (Berk, 1983). However, NLP differs fundamentally in analytic goals (namely out-of-sample prediction for NLP versus parameter inference for hypothesis testing in social science) that bring about NLP-specific situations: biases in word embeddings, annotator labels, or predicting over-amplified demographics.

specific criterion for a particular population, and is found to give systematically different predictions for subgroups of this population who are in fact identical on that specific criterion.[3]

We generalize Swinton's definition in two ways: First, to align notation with standard supervised modeling, we say there are both Y (a random variable representing the "true" values of an outcome) and $\hat{Y}$ (a random variable representing the predictions. Next, we allow the concept to apply to differences associated with continuously-valued human attributes rather than simply discrete subgroups of people.[4] Below, we define two types of measurable systematic differences (i.e. "disparities"): (1) a systematic difference between Y and $\hat{Y}$ (*outcome disparity*) and (2) a difference in error ($\epsilon = |Y - \hat{Y}|$ *error disparity*, both as a function of a given human attribute, A.

Outcome disparity. Formally, we say an *outcome disparity* exists for outcome, Y, a domain D (with values *source* or *target*), and with respect to attribute, A, when the distribution of the predicted outcome ($Q(\hat{Y}_D|A_D)$) is dissimilar to a given theoretical *ideal distribution* ($P(Y_D|A_D)$):

$$Q(\hat{Y}_D|A_D) \nsim P(Y_D|A_D)$$

The *ideal distribution* is specific to the target application. Our framework allows researchers to use their own criteria to determine this distribution. However, the task of doing so may be nontrivial. First, the current distribution within a population may not be accessible. Even when it is, it may not be what most consider the ideal distribution (e.g., the distribution of gender in computer science and the associated disparity of NLP models attributing male pronouns to computer scientists more frequently (Hovy, 2015)). Second, it may be difficult to come to an agreed-upon ideal distribution from a moral or ethical perspective. In such a case, it may be helpful to use an ideal "direction," rather than specifying a specific distribution (e.g., moving toward a uniform distribution of

---

[3]We have substituted "test" with "predictive model".

[4]"Attributes" include both continuously valued user-level variables, like age, personality on a 7-point scale, etc. (also referred to as "dimensional" or "factors"), and discrete categories like membership in an ethnic group. Psychological research suggests that people are better represented by continuously valued scores, where possible, than discrete categories (Baumeister et al., 2007; Widiger and Samuel, 2005; McCrae and Costa Jr., 1989). In NLP, Lynn et al. (2017) shows benefits from treating user-level attributes as continuously when integrating into NLP models.

pronouns associated with computer science). Our framework should enable its users to apply evolving standards and norms across NLP's many application contexts.

A **prototypical example of *outcome disparity*** is gender disparity in image captions. Zhao et al. (2017) and Hendricks et al. (2018) demonstrate a systematic difference with respect to gender in the outcome of the model, $\hat{Y}$ even when taking the source distribution as an ideal target distribution: $Q(\hat{Y}_{target}|gender) \nsim Q(Y_{target}|gender) \sim Q(Y_{source}|gender)$. As a result, captions overpredict females in images with ovens and males in images with snowboards.

Error disparity. We say there is an *error disparity* when model predictions have larger error for individuals with a given user attribute (or range of attributes in the case of continuously-valued attributes). Formally, the error of a predicted distribution is

$$\epsilon_D = |Y_D - \hat{Y}_D|$$

If there is a difference in ϵ_D over at least two different values of an attribute A (assuming they have been adequately sampled to establish a distribution of ϵ_D) then there is an error disparity:

$$Q(\epsilon_D|A_i) \nsim Q(\epsilon_D|A_j)$$

In other words, the error for one group might systematically differ from the error for another group, e.g., the error for green people differs from the error for blue people. Under unbiased conditions, the difference would be equal. This formulation allows us to capture both the discrete case (arguably more common in NLP, for example, in POS tagging) and the continuous case (for example, in age or income prediction).

We propose that if either of these two disparities exist in our target application, then there is a *predictive bias*. Note that predictive bias is then a property of a model *given* a specific application, rather than merely an intrinsic property of the model by itself. This definition mirrors predictive bias in standardized testing (Swinton, 1981): "a [predictive model] cannot be called biased without reference to a specific prediction situation; thus, the same instrument may be biased in one application, but unbiased in another."

A **prototypical example of *error disparity*** is the "Wall Street Journal Effect" – a systematic

difference in error as a function of demographics, first documented by Hovy and Søgaard (2015). In theory, POS tagging errors increase the further an author's demographic attributes differ from the average WSJ author of the 1980s and 1990s (on whom many POS taggers were trained – a selection bias, discussed next). Work by Sap et al. (2019) shows error disparity from a different origin, namely unfairness in hate speech detection. They find that annotators for hate speech on social media make more mistakes on posts of black individuals. Contrary to the case above, the disparity is not necessarily due to a difference between author and annotator population (a selection bias). Instead, the label disparity stems from annotators failing to account for the authors' racial background and sociolinguistic norms.

Source and Target Populations. An important assumption of our framework is that disparities are dependent on the population for which the model will be applied. This assumption is reflected in distinguishing a separate "target population" from the "source population" on which the model was trained. In cross-validation over random folds, models are trained and tested over the same population. However, in practice, models are often applied to novel data that may originate from a different population of people. In other words, the disparity may exist as a model property for one application, but not for another.

Quantifying disparity. Given the definitions of the two types of disparities, we can quantify bias with well-established measures of distributional divergence or deviance. Specifically, we suggest the Log-likelihood ratio as a central metric:

$$D(Y, \hat{Y}|A) = 2(log(p(Y|A)) - log(p(\hat{Y}|A)))$$

where $p(Y|A)$ is the specified ideal distribution (either derived empirically or theoretically) and $p(\hat{Y}|A)$ is the distribution within the data. For error disparity the ideal distribution is always the *Uniform* and $\hat{Y}$ is replaced with the error. KL divergence ($D_{KL}[P(\hat{Y}|A)P(Y|A)]$) can be used as a secondary, more scalable alternative.

Our measure above attempts to synthesize metrics others have used in works focused on specific biases. For example, the definition of outcome disparity is analogous to that used for semantic bias. Kurita et al. (2019) quantify bias in embeddings as the difference in log probability score when re-

placing words suspected to carry semantic differences ('he', 'she') with a mask:

$$log(P([Mask] = ``\langle PRON \rangle"|[Mask] \text{ is } ``\langle NOUN \rangle"))} -$$
$$log(P([Mask] = ``\langle PRON \rangle"|[Mask] \text{ is } [Mask])))$$

$\langle NOUN \rangle$ is replaced with a specific noun to check for semantic bias (e.g., an occupation), and $\langle PRON \rangle$ is an associated demographic word (e.g., "he" or "she").

3   Four Origins of Bias

But what leads to an outcome disparity or error disparity? We identify four points within the standard supervised NLP pipeline where bias may originate: (1) the training labels (*label bias*), (2) the samples used as observations — for training or testing (*selection bias*), (3) the representation of data (*semantic bias*), or (4) due to the fit method itself (*overamplification*).

Label Bias Label bias emerges when the distribution of the dependent variable in the data source diverges substantially from the ideal distribution:

$$Q(Y_s|A_s) \nsim P(Y_s|A_s)$$

Here, the labels themselves are erroneous concerning the demographic attribute of interest (as compared to the source distribution). Sometimes, this bias is due to a non-representative group of annotators (Joseph et al., 2017). In other cases, it may be due to a lack of domain expertise (Plank et al., 2014), or due to preconceived notions and stereotypes held by the annotators (Sap et al., 2019).

Selection bias. Selection bias emerges due to non-representative observations. I.e., when the users generating the training (source) observations differ from the user distribution of the target, where the model will be applied. Selection bias (sometimes also referred to as *sample bias*) has long been a concern in the social sciences. At this point, testing for such a bias is a fundamental consideration in study design (Berk, 1983; Culotta, 2014). Non-representative data is the origin for selection bias.

Within NLP, some of the first works to note demographic biases were due to a selection bias (Hovy and Søgaard, 2015; Jørgensen et al., 2015). A prominent example is the so-called "Wall Street Journal effect", where syntactic parsers and part-of-speech taggers are most accurate over language written by middle-aged white men. The effect occurs because

this group happened to be the predominant authors' demographics of the WSJ articles, which are traditionally used to train syntactic models (Garimella et al., 2019). The same effect was reported for language identification difficulties for African-American Vernacular English (Blodgett and O'Connor, 2017; Jurgens et al., 2017).

The predicted output is dissimilar from the ideal distribution, leading, for example, to lower accuracy for a given demographic, since the source did not reflect the *ideal distribution*. We say that the distribution of human attribute, A, within the source data, s, is dissimilar to the distribution of A within the target data, t:

$$Q(A_s) \nsim P(A_t)$$

Selection bias has several peculiarities. First, it is dependent on the *ideal distribution* of the target population, so a model may have selection bias for one application (and its associated target population), but not for another. Also, consider that either the source features (X_s) or source labels (Y_s) may be **non-representative**. In many situations, the distributions for the features and labels are the same. However, there are some cases where they diverge. For example, when using features from age-biased tweets, but labels from non-biased census surveys. In such cases, we need to take multiple analysis levels into account: corrections can be applied to user features as they are aggregated to communities (Almodaresi et al., 2017). The consequences could be both *outcome* and *error disparity*.

One of the challenges in addressing selection bias is that we can not know *a priori* what sort of (demographic) attribute will be important to control. Age and gender are well-studied, but others might be less obvious. We might someday realize that a formerly innocuous attribute (say, handedness) turns out to be relevant for selection biases. This problem is known as The Known and Unknown Unknowns.

> *As we know, there are known knowns: there are things we know we know. We also know there are known unknowns: that is to say, we know there are some things we do not know. But there are also unknown unknowns: the ones we don't know we don't know.*
> — Donald Rumsfeld

| | | ANNOTATION | |
		incorrect	**correct**
SAMPLE	**not-repr.**	selection bias, label bias	selection bias
	repr.	label bias	no bias

Table 1: Interaction between selection and label bias under different conditions for sample representativeness and annotation quality

We will see later how better documentation can help future researchers address this problem.

Overamplification. Another source of bias can occur even when there is no label or selection bias. In *overamplification*, a model relies on a small difference between human attributes with respect to the objective (even an acceptable difference matching the ideal distribution), but amplifies this difference to be much more pronounced in the predicted outcomes. The origins of overamplification are during learning itself. The model learns to pick up on imperfect evidence for the outcome, which brings out the bias.

Formally, in *overamplification* the predicted distribution ($Q(\hat{Y}_s|A_s)$) is dissimilar to the source training distribution ($Q(Y_s|A_s)$) with respect to a human attribute, A. The predicted distribution is therefore also dissimilar to the target *ideal distribution*:

$$Q(\hat{Y}_s|A_s) \nsim Q(Y_s|A_s) \sim P(Y_t|A_t)$$

For example, Yatskar et al. (2016) found that in the imSitu image captioning data set, 58% of captions involving a person in a kitchen mention women. However, standard models trained on such data end up predicting people depicted in kitchens as women 63% of the time (Zhao et al., 2017). In other words, an error in generating a gender reference within the text (e.g., "*A* [**woman** ∥ **man**] *standing next to a counter-top*") males an incorrect female reference much more common.

The occurrence of overamplification in the absence of other biases is an important motivation for countermeasures. It does not require bias on the part of the annotator, data collector, or even the programmer/data analyst (though it can escalate existing biases and the models' statistical discrimination along a demographic dimension). In particular, it extends countermeasures beyond the point some authors have made, that they are

merely cosmetic and do not address the underlying cause: biased language in society (Gonen and Goldberg, 2019).

Semantic bias. Embeddings (i.e., vectors representing the meaning of words or phrases) have become a mainstay of modern NLP, providing more flexible representations that feed both traditional and deep learning models. However, these representations often contain unintended or undesirable associations and societal stereotypes (e.g., connecting medical doctors more frequently to male pronouns than female pronouns, see Bolukbasi et al. (2016); Caliskan et al. (2017)). We adopt the term used for this phenomenon by others, "semantic bias".

Formally, we attribute semantic bias to the parameters of the embedding model (θ_{emb}). Semantic bias is a unique case since it indirectly affects both *outcome disparity* and *error disparity* by creating other biases, such as overamplification (Yatskar et al., 2016; Zhao et al., 2017) or diverging words associations within embeddings or language models (Bolukbasi et al., 2016; Rudinger et al., 2018). However, we distinguish it from the other biases, since the population does not have to be people, but rather words in contexts that yield non-ideal associations. For example, the issue is not (only) that a particular gender authors more of the training data for the embeddings. Instead, that gendered pronouns are mentioned alongside occupations according to a non-ideal distribution (e.g., texts talk more about male doctors and female nurses than vice versa). Furthermore, pretrained embeddings are often used without access to the original data (or the resources to process it). We thus suggest that embedding models themselves are a distinct source of bias within NLP predictive pipelines.

They have consequently received increased attention, with dedicated sessions at NAACL and ACL 2019. As an example, Kurita et al. (2019) quantify human-like bias in BERT. Using the Gender Pronoun Resolution (GPR) task, they find that, even after balancing the data set, the model predicts no female pronouns with high probability. Semantic bias is also of broad interest to the social sciences as a diagnostic tool (see Section A). However, their inclusion in our framework is not for reasons of social scientific diagnostics, but rather to guide mindful researchers where to look for problems.

Multiple Biases. Biases occur not only in isolation, but they also compound to increase their effects. Label and selection bias can – and often *do* – interact, so it can be challenging to distinguish them. Table 1 shows the different conditions to understand the boundaries of one or another.

Consider the case where a researcher chooses to balance a sentiment data set for a user attribute, e.g., age. This decision can directly impact the label distribution of the target variable. E.g., because the positive label is over-represented in a minority age group. Models learn to exploit this confounding correlation between age and label prevalence and magnify it even more. The resulting model may be useless, as it only captures the distribution in the synthetic data sample. We see this situation in early work on using social media data to predict mental health conditions. Models to distinguish PTSD from depression turned out to mainly capture the differences in user age and gender, rather than language reflecting the actual conditions (Preoţiuc-Pietro et al., 2015).

3.1 Other Bias Definitions and Frameworks

While this is the first attempt at a comprehensive conceptual framework for bias in NLP, alternative frameworks exist, both in other fields and based on more qualitative definitions. Friedler et al. (2016) define bias as unfairness in algorithms. They specify the idea of a "construct" space, which captures the latent features in the data that help predict the right outcomes. They suggest that finding those latent variables would also enable us to produce the right outcomes. Hovy and Spruit (2016) take a broader scope on bias based on ethics in new technologies. They list three qualitative sources (data, modeling, and research design), and suggest three corresponding types of biases: demographic bias, overgeneralization, and topic exposure. Suresh and Guttag (2019) propose a qualitative framework for bias in machine learning, defining bias as a "potential harmful property of the data". They categorize bias into historical bias, representation bias, measurement bias, and evaluation bias. Glymour and Herington (2019) classify algorithmic bias, in general, into four different categories, depending on the causal conditional dependencies to which it is sensitive: procedural bias, outcome bias, behavior-relative error bias, and score-relative error bias. Corbett-Davies and Goel (2018) propose statistical limitations of the three prominent definitions of fair-

ness (anti-classification, classification parity, and calibration), enabling researchers to develop fairer machine learning algorithms.

Our framework focuses on NLP, but it follows Glymour and Herington (2019) in providing probabilistic based definitions of bias. It incorporates and formalizes the above to varying degrees.

In social sciences, bias definitions often relate to the ability to test causal hypotheses. Hernán et al. (2004) propose a common structure for various types of selection bias. They define bias as the difference between a variable and the outcome, and the causal effect of a variable on the outcome. E.g., when the causal risk ratio (CRR) differs from associational risk ratio (ARR). Similarly, Baker et al. (2013) define bias as uncontrolled covariates or "disturbing variables" that are related to measures of interest.

Others provide definitions restricted to particular applications. For example, Caliskan et al. (2017) propose the Word-Embedding Association Test (WEAT). It quantifies semantic bias based on the distance between words with demographic associations in the embedding space. The previously mentioned work by Kurita et al. (2019) and Sweeney and Najafian (2019) extend such measures. Similarly, Romanov et al. (2019) define bias based on the correlation between the embeddings of human attributes with the difference in the True Positive rates between human traits. This approach is reflective of an error disparity.

Our framework encompasses bias-related work in the social sciences. Please see the supplement in A.1 for a brief overview.

4 Countermeasures

We group proposed countermeasures based on the origin(s) on which they act.

Label Bias. There are several ways to address label bias, typically by controlling for biases of the annotators (Pavlick et al., 2014). Disagreement between annotators has long been an active research area in NLP, with various approaches to measure and quantify disagreement through inter-annotator agreement (IAA) scores to remove outliers (Artstein and Poesio, 2008). Lately, there has been more of an emphasis on embracing variation through the use of Bayesian annotation models (Hovy et al., 2013; Passonneau and Carpenter, 2014; Paun et al., 2018). These models arrive at a much less biased estimate for the final label than

majority voting, by attaching confidence scores to each annotator, and reweighting them through that method. Other approaches have explored harnessing the inherent disagreement among annotators to guide the training process (Plank et al., 2014). By weighting updates by the amount of disagreement on the labels, this method prevents bias towards any one label. The weighted updates act as a regularizer during training, which might also help prevent overamplification. If annotators behave in predictable ways to produce artifacts (i.e., always add "not" to form a contradiction), we can train a model on such biased features and use it in ensemble learning (Clark et al., 2019). Hays et al. (2015) attempt to make Web studies equivalent to representative focus group panels. They give an overview of probabilistic and non-probabilistic approaches to generate the Internet panels that contribute to the data generation. Along with the six demographic attributes (age, gender, race/ethnicity, education, marital status, and income), they use poststratification to reduce the bias (some of these methods cross into addressing selection bias).

Selection bias. The primary source for selection bias is the mismatch between the sample distribution and the ideal distribution. Consequently, any countermeasures need to re-align the two distributions to minimize this mismatch.

The easiest way to address the mismatch is to re-stratify the data to more closely match the ideal distribution. However, this often involves downsampling an overly represented class, which reduces the number of available instances. Mohammady and Culotta (2014) use a stratified sampling technique to reduce the selection bias in the data. Almeida et al. (2015) use demographic user attributes, including age, gender, and social status, to predict the election results in six different cities of Brazil. They use stratified sampling on all the resulting groups to reduce selection bias.

Rather than re-sampling, others use *reweighting* or *poststratifying* to reduce selection bias. Culotta (2014) estimates county-level health statistics based on social media data. He shows we can stratify based on external socio-demographic data about a community's composition (e.g., gender and race). Park et al. (2006) estimate statewise public opinions using the National Surveys corpus. To reduce bias, they use various socioeconomic and demographic attributes (state of res-

idence, sex, ethnicity, age, and education level) in a multilevel logistic regression. Choy et al. (2011) and Choy et al. (2012) also use race and gender as features for reweighting in predicting the results of the Singapore and US presidential elections. Baker et al. (2013) study how selection bias manifests in inferences for a larger population, and how to avoid it. Apart from the basic demographic attributes, they also consider attitudinal and behavioral attributes for the task. They suggest using reweighting, ranking reweighting or propensity score adjustment, and sample-matching techniques to reduce selection bias.

Others have suggested combinations of these approaches. Hernán et al. (2004), propose Directed Acyclic graphs for various heterogeneous types of selection bias, and suggest using stratified sampling, regression adjustment, or inverse probability weighting to avoid the bias in the data. Zagheni and Weber (2015), study the use of Internet Data for demographic studies and propose two approaches to reduce the selection bias in their task. If the ground truth is available, they adjust selection bias based on the calibration of a stochastic microsimulation. If unavailable, they suggest using a difference-in-differences technique to find out trends on the Web.

Zmigrod et al. (2019) show that gender-based selection bias could be addressed by data augmentation, i.e., by adding slightly altered examples to the data. This addition addresses selection bias originating in the features (X_{source}), so that the model is fit on a more gender-representative sample. Their approach is similar to the reweighting of poll data based on demographics, which can be applied more directly to tweet-based population surveillance (see our last case study, A.2).

Li et al. (2018) introduce a model-based countermeasure. They use an adversarial multitask-learning setup to model demographic attributes as auxiliary tasks explicitly. By reversing the gradient for those tasks during backpropagation, they effectively force the model to ignore confounding signals associated with the demographic attributes. Apart from improving overall performance across demographics, they show that it also protects user privacy. The findings from Elazar and Goldberg (2018), however, suggest that even with adversarial training, internal representations still retain traces of demographic information.

Overamplification. In its simplest form, overamplification of inherent bias by the model can be corrected by downweighting the biased instances in the sample, to discourage the model from exaggerating the effects.

A common approach involves using synthetic matched distributions. To address gender bias in neural network approaches to coreference resolution Rudinger et al. (2018); Zhao et al. (2018) suggest matching the label distributions in the data, and training the model on the new data set. They swap male and female instances and merge them with the original data set for training. In the same vein, Webster et al. (2018) provide a gender-balanced training corpus for coreference resolution. Based on the first two corpora, Stanovsky et al. (2019) introduce a bias evaluation for machine translation, showing that most systems overamplify gender bias (see also Prates et al. (2018)). Hovy et al. (2020) show that this overamplification consistently makes translations sound older and more male than the original authors.

Several authors have suggested it is essential for language to be understood within the context of the author and their social environment Jurgens (2013); Danescu-Niculescu-Mizil et al. (2013); Hovy (2018); Yang et al. (2019). Considering the author demographics improves the accuracy of text classifiers Volkova et al. (2013); Hovy (2015); Lynn et al. (2017), and in turn, could lead to decreased error disparity.

Semantic bias. Countermeasures for semantic bias in embeddings typically attempt to adjust the parameters of the embedding model to reflect a target distribution more accurately. Because all of the above techniques can be applied for model fitting, here we highlight techniques that are more specific to addressing bias in embeddings.

Bolukbasi et al. (2016) suggest that techniques to de-bias embeddings can be classified into two approaches: hard de-biasing (completely removes bias) and soft de-biasing (partially removes bias avoiding side effects). Romanov et al. (2019) generalize this work to a multi-class setting, exploring methods to mitigate bias in an occupation classification task. They reduce the correlation between the occupation of people and the word embedding of their names, and manage to simultaneously reduce race and gender biases without reducing the classifier's performance. Manzini et al. (2019), identify the bias subspace using principal compo-

nent analysis and remove the biased components using hard Neutralize and Equalize de-biasing and soft biasing methods proposed by Bolukbasi et al. (2016). The above examples evaluate success through the semantic analogy task (Mikolov et al., 2013), a method whose informativeness has since been questioned, though (Nissim et al., 2019). For a dedicated overview of semantic de-biasing techniques see Lauscher et al. (2020).

Social-Level Mitigation. Several initiatives propose standardized documentation to trace potential biases, and to ultimately mitigate them. *Data Statements* Bender and Friedman (2018) suggest clearly disclosing data selection, annotation, and curation processes explicitly and transparently. Similarly, Gebru et al. (2018) suggest *Datasheets* to cover the lifecycle of data including "motivation for dataset creation; dataset composition; data collection process; data preprocessing; dataset distribution; dataset maintenance; and legal and ethical considerations". Mitchell et al. (2019) extend this idea to include model specifications and performance details on different user groups. Hitti et al. (2019) propose a taxonomy for assessing the gender bias of a data set. While these steps do not directly mitigate bias, they can encourage researchers to identify and communicate sources of label or selection bias. Such documentation, combined with a conceptual framework to guide specific mitigation techniques, acts as an essential mitigation technique at the level of the research community.

See Appendix A.2 for case studies outlining various types of bias in several NLP tasks.

5 Conclusion

We present a comprehensive overview of the recent literature on predictive bias in NLP. Based on this survey, we develop a unifying conceptual framework to describe bias sources and their effects (rather than *just* their effects). This framework allows us to group and compare works on countermeasures. Rather than giving the impression that bias is a growing problem, we would like to point out that bias is not necessarily something gone awry, but rather something nearly inevitable in statistical models. We do, however, stress that we need to acknowledge and address bias with proactive measures. Having a formal framework of the causes can help us achieve this.

We would like to leave the reader with these main points: (1) every predictive model with errors is bound to have disparities over human attributes (even those not directly integrating human attributes); (2) disparities can result from a variety of origins — the embedding model, the feature sample, the fitting process, and the outcome sample — within the standard predictive pipeline; (3) selection of "protected attributes" (or human attributes along which to avoid biases) is necessary for measuring bias, and often helpful for mitigating bias and increasing the generalization ability of the models.

We see this paper as a step toward a unified understanding of bias in NLP. We hope it inspires further work in both identifying and countering bias, as well as conceptually and mathematically defining bias in NLP.

Framework Application Steps (TL;DR)

1. Specify *target* population and an *ideal distribution* of the attribute (A) to be investigated for bias; Consult datasheets and data statements[5] if available for the model *source*;
2. If *outcome disparity* or *error disparity*, check for potential origins:
   (a) if *label bias*: use post-stratification or retrain annotators.
   (b) if *selection bias*: use stratified sampling to match source to target populations, or use post-stratification, re-weighting techniques.
   (c) if *overamplification*: synthetically match distributions or add outcome disparity to cost function.
   (d) if *semantic bias*: retrain or retrofit embeddings considering approaches above, but with attributed (e.g., gendered) words (rather than people) as the population.

---

[5] (Gebru et al., 2018; Bender and Friedman, 2018)

References

Jussara M Almeida, Gisele L Pappa, et al. 2015. Twitter population sample bias and its impact on predictive outcomes: A case study on elections. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1254–1261. ACM.

Fatemeh Almodaresi, Lyle Ungar, Vivek Kulkarni, Mohsen Zakeri, Salvatore Giorgi, and H. Andrew Schwartz. 2017. On the distribution of lexical features at multiple levels of analysis. In *The 55th Annual Meeting of the Association for Computational Linguistics*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Reg Baker, J Michael Brick, Nancy A Bates, Mike Battaglia, Mick P Couper, Jill A Dever, Krista J Gile, and Roger Tourangeau. 2013. Summary report of the aapor task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1(2):90–143.

Roy F Baumeister, Kathleen D Vohs, and David C Funder. 2007. Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, 2(4):396–403.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 152–162.

Richard A Berk. 1983. An introduction to sample selection bias in sociological data. *American Sociological Review*, pages 386–398.

Sudeep Bhatia. 2017. Associative judgment and vector space semantics. *Psychological review*, 124(1):1.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media African-American English. *arXiv preprint arXiv:1707.00061*.

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. In *Advances in neural information processing systems*, pages 4349–4357.

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

Murphy Choy, Michelle Cheong, Ma Nang Laik, and Koo Ping Shung. 2012. Us presidential election 2012 prediction using census corrected twitter model. *arXiv preprint arXiv:1211.0938*.

Murphy Choy, Michelle LF Cheong, Ma Nang Laik, and Koo Ping Shung. 2011. A sentiment analysis of singapore presidential election 2011 using twitter data with census correction. *arXiv preprint arXiv:1108.5520*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Donâ ĂŹt take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4060–4073.

Maximin Coavoux, Shashi Narayan, and Shay B Cohen. 2018. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *CLPsych@ HLT-NAACL*, pages 31–39.

Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Fintan Costello and Paul Watts. 2014. Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological review*, 121(3):463.

Mick P Couper. 2013. Is the sky falling? new technology, changing media, and the future of surveys. In *Survey Research Methods*, volume 7, pages 145–156.

Aron Culotta. 2014. Reducing sampling bias in social media data for county health inference. In *Joint Statistical Meetings Proceedings*, pages 1–12.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. corr abs/1609.07236 (2016).

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. In *Proceedings of the 5 th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden.

Salvatore Giorgi, Veronica Lynn, Keshav Gupta, Sandra Matz, Lyle Ungar, and H.A. Schwartz. 2019. Correcting sociodemographic selection biases for population prediction. *ArXiv*.

Bruce Glymour and Jonathan Herington. 2019. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 269–278, New York, NY, USA. ACM.

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 *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.

Ron D Hays, Honghu Liu, and Arie Kapteyn. 2015. Use of internet panels to conduct surveys. *Behavior research methods*, 47(3):685–690.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. 2004. A structural approach to selection bias. *Epidemiology*, pages 615–625.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy. Association for Computational Linguistics.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 752–762.

Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People'âĂŹs Opinions, Personality, and Emotions in Social Media*, pages 42–49.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. Can you translate that into man? commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, WA, USA. Association for Computational Linguistics.

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)*, volume 2, pages 483–488.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 591–598.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 264–271.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18.

Kenneth Joseph, Lisa Friedland, William Hobbs, Oren Tsur, and David Lazer. 2017. Constance: Modeling annotation contexts to improve stance classification. *arXiv preprint arXiv:1708.06309*.

David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Seventh International AAAI Conference on Weblogs and Social Media*.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57.

ML Kern, G Park, JC Eichstaedt, HA Schwartz, M Sap, LK Smith, and LH Ungar. 2016. Gaining insights from social media language: Methodologies and challenges. *Psychological methods*, 21(4):507–525.

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.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2018. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2018. The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 25–30.

Veronica E. Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered nlp with user-factor adaptation. In *Empirical Methods in Natural Language Processing*.

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*.

Robert R. McCrae and Paul T. Costa Jr. 1989. Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality. *Journal of Personality*, 57(1):17–40.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229. ACM.

Ehsan Mohammady and Aron Culotta. 2014. Using county demographics to infer attributes of twitter users. In *Proceedings of the joint workshop on social dynamics and personal attributes in social media*, pages 7–16.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2019. Fair is better than sensational: Man is to doctor as woman is to doctor. *arXiv preprint arXiv:1905.09866*.

David K Park, Andrew Gelman, and Joseph Bafumi. 2006. State-level opinions from national surveys: Poststratification using multilevel logistic regression. *Public opinion in state politics*, pages 209–28.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

James W Pennebaker. 2011. The secret life of pronouns. *New Scientist*, 211(2828):42–45.

James W. Pennebaker and Lori D. Stone. 2003. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2):291.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator

agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2018. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, pages 1–19.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 21–30.

Daniel Preotiuc-Pietro, Maarten Sap, H Andrew Schwartz, and Lyle Ungar. 2015. Mental illness detection at the world well-being project for the clpsych 2015 shared task. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, NAACL*.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. 2019. Machine behaviour. *Nature*, 568(7753):477.

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. *arXiv preprint arXiv:1904.05233*.

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)*, volume 2, pages 8–14.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, Lyle Ungar, et al. 2013. Characterizing geographic variation in well-being using tweets. In *Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*.

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.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

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

Spencer S Swinton. 1981. Predictive bias in graduate admissions tests. *ETS Research Report Series*, 1981(1):i–53.

Peter Trudgill. 2000. *Sociolinguistics: An introduction to language and society*. Penguin UK.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827.

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.

Thomas A. Widiger and Douglas B. Samuel. 2005. Diagnostic categories or dimensions? A question for the Diagnostic and Statistical Manual of Mental Disorders—Fifth Edition. *Journal of Abnormal Psychology*, 114(4):494.

Diyi Yang, Robert E Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542.

Emilio Zagheni and Ingmar Weber. 2015. Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1):13–25.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1651–1661.

A Appendices

A.1 Related Work in Other Fields

We survey the literature of adjacent fields and outline different streams related to our framework. These examples illustrate the ubiquity and complexity of bias, and highlight its understanding in different disciplines over time.

Bias became a crucial topic in social science following the seminal work of Tversky and Kahneman, showing that human thinking was subject to systematic errors (Tversky and Kahneman, 1973). Human logic was seemingly separate from the principles of probability calculus. "Bias" here is interpreted as the result of psychological heuristics, i.e., mental "shortcuts" to help us react faster to situations. While many of these heuristics can be useful in critical situations, their indiscriminate application in everyday life can have adverse effects and cause bias. This line of work has spawned an entire field of study in psychology (decision making).

The focus of Tversky and Kahneman (1973) (and a whole field of decision making that followed) was human behavior. Still, the same basic principles of systematic differences in decision making apply to machines as well. However, algorithms also provide systematic ways to reduce bias, and some see the mitigation of bias in algorithm decisions as a potential opportunity to move the needle positively (Kleinberg et al., 2018). Thus, we can apply frameworks of contemporaries in human behavior to machines (Rahwan et al., 2019), and perhaps benefit from a more scalable experimentation process. Costello and Watts (2014) studies human judgment under uncertain conditions, and proposes that we *can* algorithmically account for observed human bias, provided there is sufficient random noise in the probabilistic model. This view suggests bias within the model itself, what we have called *overamplification*.

Still, most works on bias in decision making assume working with unbiased data, even though social science has long battled selection bias. Most commonly, data selection is heavily skewed towards the students found on western university campuses (Henrich et al., 2010). Attempts to remedy selection bias in a scalable fashion use online populations, which in turn are skewed by unequal access to the Internet, but which we can mitigate through reweighting schemes (Couper, 2013).

In some cases, algorithmic bias has helped understand society better. For example, *semantic bias* in word embeddings has been leveraged to track trends in societal attitudes concerning gender roles and ethnic stereotypes. Garg et al. (2018); Kozlowski et al. (2018) measure the distance between certain sets of words in different decades to track this change. This use of biased embeddings illustrates an interesting distinction between *normative* and *descriptive* ethics. When used in predictive models, semantic bias is something to be avoided (Bolukbasi et al., 2016). I.e., it is *normatively wrong* for many applications (e.g., we ideally would want all genders or ethnicities equally associated with all jobs). However, the works by Garg et al. (2018) and Kozlowski et al. (2018) show that it is precisely this bias of word embeddings that reflects societal attitudes. Here, the presence of bias is *descriptively correct*. Similarly, Bhatia (2017) uses this property of word embeddings to measure people's psychological biases and attitudes towards making individual decisions.

A.2 Discussion: Example Case Studies

Part of Speech Taggers and Parsing. The works by Hovy and Søgaard (2015); Jørgensen et al. (2015) outline the effect of selection bias on syntactic tools. The language of demographic groups systematically differs from each other for syntactic attributes. Therefore, models trained on samples whose demographic composition (e.g., age and ethnicity) differs from the target perform significantly worse. Within the predictive bias framework, the consequence of this selection bias is an *error disparity* − $Q(\epsilon_{D=general}|A = age, ethnicity) \not\sim Uniform$, the error of the model across a general domain (D) is not uniform with respect to attributes (A) age and ethnicity. Li et al. (2018) shows that this consequence of selection bias can be addressed by adversarial learning, removing the age gap and significantly reducing the performance difference between ethnolects (even if it was not trained with that objective). Garimella et al. (2019) quantifies this bias further by studying the effect of different gender compositions of the training data on tagging and parsing, supporting the claim that debiased samples benefit performance.

Image Captions. Hendricks et al. (2018) shows the presence of gender bias in image captioning, overamplifying differences present in the train-

ing data. Prior work focused on context (e.g., it is easier to predict "mouse" when there is a computer present). This bias manifests in ignoring people present in the image. The gender bias is not only influenced by the images, but also by biased language models. The primary consequence is an *outcome disparity* − $Q(\hat{Y}_D|gender) \not\approx P(Y_D|gender)$, the distribution of outcomes (i.e. caption words and phrases) produced from the model $Q(\hat{Y}_D|gender)$ over-selects particular phrases beyond the distribution observed in reality: (i.e. $P(Y_D|gender)$; this is true even when the source and target are the same: $D = source = target$).

To overcome the bias and to increase performance, Hendricks et al. (2018) introduce an equalizer model with two loss-terms: Appearance Confusion Loss (ACL) and Confident Loss (Conf). ACL increases the gender confusion when gender information is not present in the image, making it difficult to predict an accurately gendered word. Confident loss increases the confidence of the predicted gendered word when gender information *is* present in the image. Both loss terms have the effect of decreasing the difference between $Q(\hat{Y}_D|gender)$ and $P(\hat{Y}_D|gender)$. In the end, the Equalizer model performs better in predicting a woman while still misclassifying a man as a woman, but decreasing *error disparity* overall.

Sentiment Analysis. Kiritchenko and Mohammad (2018) show the issues of both semantic bias and overamplification. They assess scoring differences in 219 sentiment analysis systems by switching out names and pronouns. (They switch between male and female pronouns, and between prototypical white and black American first names based on name registers.) The results show that male pronouns are associated with higher scores for negative polarity, and prototypical black names with higher scores for negative emotions. The consequence of the semantic bias and overamplification are outcome disparities: $Q(\hat{Y}_D|gender) \not\approx P(Y_D|gender)$ and $Q(\hat{Y}_D|race) \not\approx P(Y_D|race)$. This finding again demonstrates a case of descriptive vs. normative ethics. We could argue that because aggression is more often associated with male protagonists, the models reflect a descriptively correct (if morally objectionable) societal fact. However, if the model score changes based on ethnicity, the difference likely reflects (and amplifies) societal ethnic stereotypes. Those stereo-

types, though, are both normatively and descriptively wrong.

Differential Diagnosis in Mental Health. In the clinical community, differentiating a subject with post-traumatic stress disorder (PTSD) from someone with depression is known to be difficult. It was, therefore, surprising when early work on this task produced AUCs greater than 0.85 (this and similar tasks were part of the CLPsych2015 Shared task; (Coppersmith et al., 2015)). Labels of depression and PTSD had been automatically derived from a convenience sample of individuals[6] who had publicly stated their diagnosis in their profile. The task included a 50/50 split from each category. However, Preotiuc-Pietro et al. (2015) show that these classifiers primarily picked up on differences in age or gender – subjects with PTSD were more likely to be older than those with depression. While age and gender themselves are valid information for mental health diagnosis, the design yielded classifiers that predicted nearly all older individuals to have PTSD, and those younger to have depression, despite the 50/50 split. These classifiers resulted in *outcome disparity*, because older individuals were much less likely to be labeled depressed than the target population (and younger less likely for PTSD: $Q(\hat{Y}_D|A = age) \not\approx P(Y_D|A = age)$). In the end, the task organizers mitigated the issue by using matched controls – adding another 50% samples for each class such that the age and gender distributions of both groups matched. Recently, Benton et al. (2017) showed that accounting for demographic attributes in the model could leverage this correlation while controlling for the confounds.

Assessing Demographic Variance in Language. A final case study in applying our framework demonstrates how inferring user demographics can mitigate bias. Consider the task of producing population measurements from readily available (but biased) community corpora. E.g., assessing representative US county life satisfaction from tweets (Schwartz et al., 2013). Unlike our other examples, the *outcomes* of the source training data (i.e., surveys) are expected to be representative, while the *features* come with biases. The source feature distributions with respect to human attributes are dissimilar from the ideal distribu-

---

[6] A convenience sample, a term from social science, is a set of data selected because it is available rather than designed for the given task.

tion, while the source outcomes match that target outcomes ($Q(X_{source}|A) \not\sim P(X_{target}|A)$ but $Q(Y_{source}|A) \sim P(Y_{target}|A)$).

In this case, the effectiveness of countermeasures preventing selection and semantic biases (for X_{source} and X_{target}) should result in increased predictive performance against a representative community outcome. Indeed, Giorgi et al. (2019) adjust the feature estimates, X, to match representative demographics and socio-economics by using inferred user attributes, and find improved predictions for the life satisfaction of a Twitter community.