# 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**
University of Bocconi
dirk.hovy@unibocconi.it

## Abstract

An increasing number of works in natural language processing have addressed the effect of bias on the predicted outcomes, introducing mitigation techniques that act on different parts of the standard NLP pipeline (data and models). However, these works have been conducted in isolation, without a unifying framework to organize efforts within the field. This leads to repetitive approaches, and puts an undue focus on the *effects* of bias, rather than on their *origins*. Research focused on bias symptoms rather than the underlying origins could limit the development of effective countermeasures. In this paper, we propose a unifying conceptualization: *the predictive bias framework for NLP*. We summarize the NLP literature and propose a general mathematical definition of predictive bias in NLP along with a conceptual framework, differentiating four main origins of biases: *label bias*, *selection bias*, *model overamplification*, and *semantic bias*. We discuss how past work has countered each bias origin. Our framework serves to guide an introductory overview of predictive bias in NLP, integrating existing work into a single structure and opening avenues for future research.

## 1 Introduction

Predictive models in NLP are sensitive to a variety of (unintended) biases throughout the development process. As a result, fitted models do not generalize well, incurring performance losses on unseen data, less reliable predictions, and socially undesirable effects by systematically under-serving or mispredicting certain user groups.

The general phenomenon of biased predictive models is perhaps not recent in NLP: 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 on socially undesirable signals during the training process. In the case of domains, these non-generalizations are often words, phrases, or senses that occur in one text type, but not another.

However, this kind of variation is not just restricted to content domains but 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. 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: biases can be seen as 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* – the distribution of labels produced by a predictive model reflect a human factor in way that diverges from a theoretically defined "desired 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 (e.g. Li et al. (2018); Elazar and Goldberg (2018); Coavoux et al. (2018)).[1] Each identifies a specific bias and countermeasure on their own 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* rather than their *origins*. While it is important to address the effects, it can leave the origin unchanged, requiring researchers to rediscover the issue over and over (Gonen and Goldberg, 2019). The "bias" discussed in one paper may therefore be quite different than that in another.[2]

A shared definition and framework of predictive bias would unify these efforts, provide a common terminology, help to identify underlying causes, and allow coordination of countermeasures (Sun et al., 2019). However, as of yet, such a general framework has 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 selected works in adjacent fields. We identify four distinct sources of bias: **selection bias**, **label bias**, **model overamplification**, and **semantic bias**. All of these can be expressed as differences between (a) a "true" or intended distribution (over users, labels, or outcomes), and (b) a distribution that is actually used or produced by the model, and they can be connected to 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, which we hope will 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. In order to ensure we are not narrowing our definition, we consider examples of bias from both within NLP and from related fields.

[1] There is an even larger body of work on *fairness* as part of the FAT* conferences, which goes beyond the scope of this paper: we are firmly focused on bias.

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

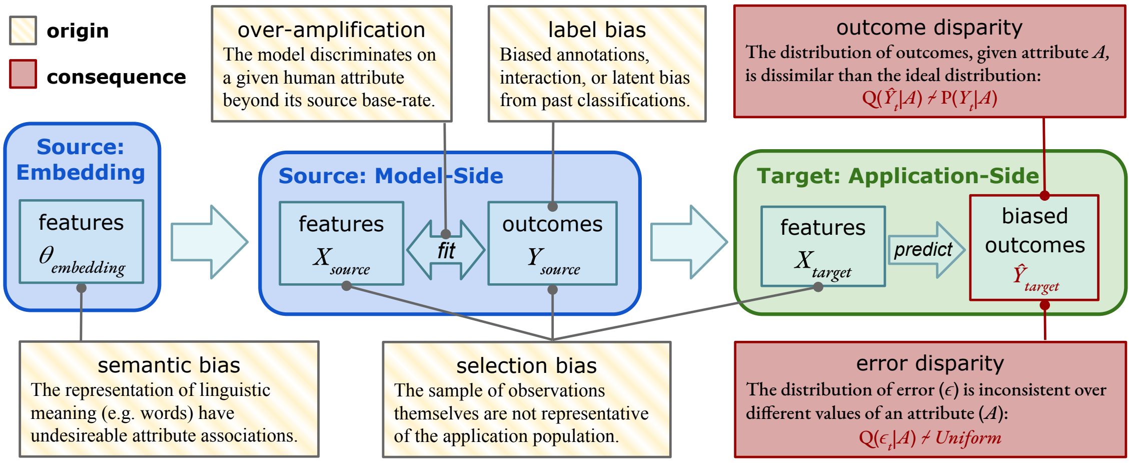
origin
consequence
over-amplification
The model discriminates on a given human attribute beyond its source base-rate.
label bias
Biased annotations, interaction, or latent bias from past classifications.
outcome disparity
The distribution of outcomes, given attribute $A$, is dissimilar than the ideal distribution:
$Q(\hat{Y}_t|A) \neq P(Y_t|A)$
Source: Embedding
features
$\theta_{embedding}$
Source: Model-Side
features
$X_{source}$
fit
outcomes
$Y_{source}$
Target: Application-Side
features
$X_{target}$
predict
biased outcomes
$\hat{Y}_{target}$
semantic bias
The representation of linguistic meaning (e.g. words) have undesireable attribute associations.
selection bias
The sample of observations themselves are not representative of the application population.
error disparity
The distribution of error ($\epsilon$) is inconsistent over different values of an attribute ($A$):
$Q(\epsilon_t|A) \neq Uniform$

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*.

**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 within our conceptual framework. Finally, (4) we present case studies where bias may unintuitively emerge or lead to downstream problems, connecting the characteristics to our 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 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 to apply to differences associated with continuously valued human *factors* rather than simply discrete subgroups of people.[4] We define two types of measurable systematic differences (i.e. "disparities") in: *outcome disparity* and *error disparity*.

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

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

[4] "Factors" is intended to be inclusive of both continuously valued user-level variables (also refereed to as "dimensional"; e.g., real-valued age) and discrete categories (e.g., 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). Lynn et al. (2017) showed moderate benefits from treating user-level factors as continuously when integrating into NLP models.

$(P(Y_D|A_D))$:

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

The *ideal distribution* is given and determined by the target application (i.e. this definition is agnostic to the moral or ethical considerations one uses to determine the *ideal distribution*).

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

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

If this difference $\epsilon_D$ is not distributed uniformly with respect to $A_D$ then there is an error disparity:

$$Q(\epsilon_D|A_D) \nsim Uniform$$

In other words, the error with respect to one group might systematically differ from the error with respect to another group, e.g., the error for green people differs from the error for blue people. Under unbiased conditions, the difference would be drawn from a uniform distribution. This formulation allows us to capture both the discrete case (which is arguably the 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 this makes bias a property of a model *given* a specific application, rather than simply an innate property of the model by itself. This 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."

**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 more focused works on specific biases. For example, from the perspective of semantic bias, Kurita et al. (2019) quantified bias in BERT as the difference in log probability score when replacing words suspected to carry semantic differences ('he', 'she') with a mask:

$$\log(P([Mask] = \text{``}\langle PRN\rangle\text{''}|[Mask]\ is\ \text{``}\langle NN\rangle\text{''})) - \log(P([Mask] = \text{``}\langle PRN\rangle\text{''}|[Mask]\ is\ [Mask])))$$

$\langle NN\rangle$ is replaced with a specific noun to check for semantic bias (e.g., an occupation) and $\langle PRN\rangle$ is a demographic associated word (e.g., "he" or "she"). Our proposed framework incorporates this approach.

## 3 Origins of Bias

But what leads to an outcome disparity or error disparity? We discuss four common 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 source diverges substantially from the ideal distribution:

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

Here, the labels themselves are erroneous with respect to the demographic factor of interest (as compared to the source distribution). In some cases this is due to having a non-representative set of language annotators (Joseph et al., 2017), while in other cases it may be due to a lack of expertise in the domain (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 – when the users generating the training (source) observations have a different distribution than that where the model is intended to be applied (target). Selection bias (sometimes also referred to as *sample bias*) has long been a concern in the social sciences to the point where considerations for such bias are now considered a fundamental consideration for study design (Berk, 1983; Culotta, 2014).

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 being 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 — which happened to be the predominant authors demographics of the WSJ articles traditionally used for training those 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).

Non-representative data is the origin for selection bias. 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*. Thus, we say that the distribution of human factor, $A$, within the source data, $s$ is dissimilar to the distribution of $A$ within the target, $t$:

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

Selection bias has several idiosyncrasies. 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**. While in many situations the distributions for the features and labels are the same, there are some where they diverge. For example, when using features from age-biased tweets, but labels from non-biased census surveys. In such cases, multiple analysis levels need to be taken 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 is is not known *a priori* what sort of (demographic) factor will be important to control. Age and gender are well-studied, but others might be less obvious. We might some day realize that a formerly innocuous factor (say, handedness) turns out to be relevant for selection biases. This is the problem of 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*

| | | 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

> *not know. But there are also unknown unknowns: the ones we don't know we don't know.*
> — Donald Rumsfeld

**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 factors with respect to the objective (even an acceptable difference matching the ideal distribution), but amplifies this difference to be much larger in the predicted outcomes. The origins of overamplification are during learning itself, and do not require bias on the part of the annotator, data collector, or even the programmer/data analyst (though it can appear when another bias exists, escalating the models' statistical discrimination along a demographic dimension). Rather, 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 factor, $A$. 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, in the imSitu image captioning data set (Yatskar et al., 2016), 58% of captions involving a person in a kitchen were found to mention women, but when standard models were trained on such data, they ended up predicting women in kitchens 63% of the time (Zhao et al., 2017). In other words, when there was an error in generating a gender reference within the text (e.g., "*A* **[woman || man]** *standing next to a counter-top*"), an incorrect female reference was much more common.

The fact that overamplification may occur in absence of other biases is important in motivating the need for countermeasures to bias. In particular, it extends countermeasures beyond the point some make that they are 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 a more flexible representation that easily feeds into deep machine learning techniques. However, many times these representations contain unintended or undesirable associations and societal stereotypes (e.g., connecting medical doctors more frequently to male pronouns than female pronouns) (Bolukbasi et al., 2016; Caliskan et al., 2017). This is semantic bias.

Formally, we attribute semantic bias to the parameters of the embedding model ($\theta_{emb}$). Semantic bias is a special case in that it indirectly effects both outcome and error disparities, by causing other biases, from overamplification (Yatskar et al., 2016; Zhao et al., 2017), to diverging words associations within embeddings or language models (Bolukbasi et al., 2016; Rudinger et al., 2018). Further, semantic bias themselves may originate from a selection bias or overamplification during fit for the embedding model itself. However, to avoid potential recursive definitions for practical benefit we suggest the embedding model itself is considered a potential source of bias within NLP predictive pipelines.

While semantic biases are of wider interest to the social sciences as a diagnostic tool (see Section 4), they have also received increased interest in NLP, with dedicated sessions at NAACL and ACL 2019. As an example, Kurita et al. (2019) quantified human-like bias in BERT. Using the Gender Pronoun Resolution (GPR) task, they found that, even after the dataset was balanced, the model predicted no female pronouns with a high probability.

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

Consider the case where a researcher chooses to balance the data set with respect to user factors, e.g., age. This might then directly impact the label distribution of the target variable (for example because it is overrepresented in a minority age group). Models will learn to exploit this confounding correlation between age and label prevalence and magnify it even more. The resulting predictions may be useless, as they only serve to capture the distribution in the synthetic data sample. This was the case for early work on using social media language to predict mental health conditions, where models that distinguished PTSD from depression were found to mostly just capture differences in age and gender of users, rather than language reflective of the conditions' actual symptoms (Preoţiuc-Pietro et al., 2015) – see Example Case Studies.

### 3.1 Other Bias Definitions and Frameworks

While we believe this is the first attempt at a comprehensive conceptual framework for bias in NLP, alternative frameworks exist from other fields more generally or based more in qualitative definitions. Friedler et al. (2016) defined bias as unfairness in algorithms. They specified 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) took a wider scope on bias, and noted three qualitative sources: data, modeling, and research design and suggested 3 different types of bias: demographic bias, overgeneralization, and topic exposure. Suresh and Guttag (2019) suggested a qualitative framework for bias in machine learning, defining bias as a "potential harmful property of the data". They categorized bias into: Historical bias, Representation bias, Measurement bias, Evaluation bias. Glymour and Herington (2019), classified algorithmic bias, in general, in 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. Our framework is more directly aligned with NLP but it follows Glymour in providing probabilistic based definitions of bias.

From the social science side, 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 outcome (i.e., when the causal risk ratio (CRR) differs from associational risk ratio (ARR). Similarly, Baker et al. (2013) defined bias as uncontrolled covariates or “disturbing variables” that are related to measures of interest.

Others provided definitions restricted to particular applications. For example, Caliskan et al. (2017) proposed the Word-Embedding Association Test (WEAT), which quantified semantic bias based on the distance between words with demographic associations in the embedding space. The previously mentioned work of Kurita et al. (2019) as well as that of Sweeney and Najafian (2019) extended such measures. Similarly, Romanov et al. (2019) defined bias based on correlation between embeddings of human attributes or human names with difference in the True Positive Rates between human attributes (reflective of an error disparity).

## 4 Related Work in Other Fields

We outline the different streams in the literature of related fields that relate to our framework. This also serves to illustrate the ubiquity and complexity of bias, and its understanding in different disciplines over time.

Bias became a key topic of social science following the seminal work of Tversky and Kahneman showing that human thinking was subject to fairly reliable systematic errors (Tversky and Kahneman, 1973). In other words, human logic was seemingly unbounded to the principles of probability calculus. “Bias” was thus 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 negative effects and cause bias.

The focus of Tversky and Kahneman (and a whole field of decision making that followed) was human behavior, but the same basic principles concerning systematic difference in decision making apply to machines. However, machines 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. For example, Costello and Watts (2014) studied the behavior of human judgment under uncertain conditions and proposed that bias could be accounted for in observed human behavior – provided there is sufficient random noise in the probabilistic model. This suggests bias within the model itself, which we term *overamplification*.

Still, most works on bias in decision making assume one is working with unbiased data, even though social science has long battled selection bias. Most commonly, 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 are skewed by unequal access to the Internet, but which can be mitigated through reweighting schemes (Couper, 2013).

In a few cases, algorithmic bias has been used to better understand society. For example, *semantic bias* emerging from word embeddings has been leveraged to track trends in societal attitudes across time with respect to gender roles and ethnic stereotypes, by measuring the distance between certain sets of words in different decades (Garg et al., 2018; Kozlowski et al., 2018). This distinction in studying 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., *normatively wrong* for many applications (e.g., for reviewing job candidates: ideally, we would want all genders or ethnicities equally associated with all jobs). However, the aforementioned works by Garg et al. (2018) and Kozlowski et al. (2018) have shown that it is precisely this property of word embeddings that reflects societal attitudes, so the presence of bias is *descriptively correct*. Similarly, Bhatia (2017) has shown that this property of word embeddings can be used to measure people's psychological biases and attitudes towards making certain decisions.

## 5 Countermeasures

Based on the predictive bias framework for NLP, we group proposed countermeasures into 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; Paun et al., 2018), which have been shown to arrive at a much less biased estimate for the final label than majority voting, by attaching confidence score to each annotator and re-weighting 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, and the weighted updates essentially act as a regularizer during training, which might also help to prevent overamplification.

Others attempt to make Web studies equivalent to representative focus group panels. Hays et al. (2015) gave an overview of probabilistic and non-probabilistic approaches to generate Internet panels, who will be participating in generating data. Along with the six demographic features used (age, gender, race/ethnicity, education, marital status, and income), they used poststratification to reduce the bias (some of these methods cross into addressing selection bias).

**Selection Bias.** The main source for selection bias is the mismatch between the sample distribution and the ideal distribution. Any countermeasures that address selection bias are therefore designed to re-align the two distributions to minimize the difference.

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, and therefore reduces the amount of available instances. Mohammady and Culotta (2014) used stratified sampling technique to reduce the selection bias in the data. Almeida et al. (2015), used demographics of users which included age, gender and social status to predict the election results in six different cities of Brazil. They used stratified sampling on all the groups formed to reduce the selection bias.

Rather than re-sampling, others weight sample instances to reduce selection bias (also known as *reweighting* or *poststratifying*). Culotta (2014) estimated the county-level health statistics based on social media data. He showed that stratification can be based on external socio-demographic (gender and race) data about a community's composition. Park et al. (2006) estimated state-wise public opinions using the National Surveys corpus. They used various socioeconomic and demographic features like state of residence, sex, ethnicity, age and education level of people within multilevel logistic regression to reduce the bias.

Choy et al. (2011) and Choy et al. (2012) also used race and gender of individuals as features for reweighting when predicting the results of Singapore and US presidential elections and vote percentage. Baker et al. (2013), studied methods by which the selection bias of samples manifests in making inferences for a larger population and how they could be avoided. Apart from the basic demographic features, attitudinal and behavioral features were also considered for the task. They suggested using reweighting, raking reweighting or propensity score adjustment and sample matching techniques to reduce the selection bias in their task.

Others have suggested combinations of these approaches. For example, Hernán et al. (2004), proposed using Directed Acyclic graphs to various heterogeneous types of selection bias, and suggested using stratified sampling, regression adjustment or inverse probability weighting to avoid the bias in the data. Zagheni and Weber (2015), studied the use of Internet Data for demographic studies and proposed two approaches to reduce the selection bias in their task. If the ground truth was available, they adjusted selection bias based on the calibration of the stochastic microsimulation. If it was not available, they suggested using a difference in differences technique to find out trends on the Web.

Recently, Zmigrod et al. (2019) showed that gender-based selection bias can be addressed by data augmentation, i.e., by adding slightly altered examples to the data. In this way, they address selection bias originating at the features ($X_{source}$) such that the model is fit to a more gender-represented sample of data. This mimics how selection bias is accounted for in reweighting poll data based on demographics, which can be applied more directly tweet-based population surveillance (see our last case study).

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

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

A common approach to avoid overamplification involves using synthetic matched distributions. To address gender bias in neural network approaches Rudinger et al. (2018); Zhao et al. (2018), suggested matching the distributions of labels in the data and train the coreference resolution model on the new dataset. In their task, they swapped the male and female instances and merged them with the original dataset for training. In the same vein, Webster et al. (2018) provided a gender-balanced training corpus for coreference resolution. Based on the first two corpora, Stanovsky et al. (2019) introduced a bias evaluation for machine translation, showing that most systems overamplify gender bias (see also Prates et al. (2018)). However, others have suggested it is important for language to be understood within the context of the author. Often the accuracy of text classifiers can be improved by considering the author demographics Hovy (2015); Lynn et al. (2017), which in turn could lead to decreased error disparity.

**Semantic Bias.** Concerning embeddings or language models themselves, countermeasures for semantic bias typically attempt to adjust the parameters of a language model to more accurately reflect a target distribution. Because all of the above techniques can be applied during language model fit, here we highlight techniques that are more specific to addressing bias in embeddings.

Bolukbasi et al. (2016) suggested that techniques to debias embeddings could be classified into two approaches: hard debiasing (completely removes bias) and soft debiasing (partially removes bias avoiding side effects). Romanov et al. (2019) generalized this work to a multi-class setting, exploring methods to mitigate bias in an occupation classification task. They reduced the correlation between the occupation of the people and the word embedding of the names of the people to reduce the bias. They were able to simultaneously reduce the race and gender biases without reducing the classifier‘s performance (True Positive rate). Manzini et al. (2019), identified the bias subspace using principal component analysis and removed the biased components using hard Neutralize and Equalize debiasing and soft biasing methods proposed by Bolukbasi et al. (2016).

**Social-Level Mitigation.** Recently, a couple initiatives proposed to standardize documentation to make clear any potential biases and ultimately mitigate them. *Data Statements* Bender and Friedman (2018) suggest clearly disclosing data selection, annotation, and curation processes explicitly and transparently. Similarly, *Datasheets* Gebru et al. (2018) are suggested 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”. Gebru et al. also noted such documentation would probably benefit from evolving over time. Mitchell et al. (2019) extended this idea to also include the model specifications and performance details on different user groups. Further, Hitti et al. (2019) proposed a taxonomy for assessing the gender bias of a dataset itself. While these steps do not directly mitigate bias, as research norms, they can encourage researchers to identify and communicate sources of label or selection bias. Combined with a conceptual framework to provide guidance on specific mitigation techniques, such documentation can be seen as an important mitigation technique acting at the level of the research community.

## 6 Discussion: Example Case Studies

**Part of Speech Taggers and Parsing.** The works by Hovy and Søgaard (2015) and Jørgensen et al. (2015) outlined the effect of selection bias on syntactic tools. Because the language of certain demographic groups systematically differs with respect to syntactic factors, models trained on biased samples that are mainly composed of a different demographic group than the target (here: age and ethnicity), perform significantly worse. Within the predictive bias framework, the consequence of this selection bias is an *error disparity* – $Q(\epsilon_{D=general}|A = age, ethnicity) \nsim Uniform$, 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) showed 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). The work by Garimella et al. (2019) has quantified 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) showed the presence of gender bias in image captioning, overamplifying differences present in the training data. Prior work has focused on context (e.g., it is easier to predict "mouse" when there is a computer present). This resulted 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* – $Q(\hat{Y}_D|gender) \nsim P(Y_D|gender)$, 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) introduced 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 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 performed better in predicting a woman, while misclassifying a man as a woman, decreasing *error disparity* overall.

**Sentiment Analysis.** The work by Kiritchenko and Mohammad (2018) shows the issues of both semantic bias and overamplification. They assessed scoring differences in 219 sentiment analysis systems when switching out names and pronouns (switching between male and female pronouns, and between prototypical white and black American first names based on name registers). The results showed that male pronouns were associated with higher scores for negative polarity, and black names with higher score for negative emotions. The consequence of the semantic bias and overamplification were outcome disparities: $Q(\hat{Y}_D|gender) \nsim P(Y_D|gender)$ and $Q(\hat{Y}_D|race) \nsim P(Y_D|race)$. This again demonstrates a case of descriptive vs. normative ethics: it could be argued that because aggression is more often associated with male protagonists, the models reflect a descriptively correct (if morally objectionable) societal fact. However, in the case where the model score changed based on ethnicity, the difference is likely due to reflecting and amplifying societal ethnic stereotypes.

**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 performing this classification from tweets 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 were automatically derived from a convenience sample of individuals[5] who had publicly stated their diagnosis in their profile, and the task was set up so that 50% were from each category. However, Preotiuc-Pietro et al. (2015) showed that these classifiers primarily picked up on differences in age and/or gender – those with PTSD were more likely to be older. While age and gender themselves are valid information for mental health diagnosis,

[5] 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.

the design using a 50/50 split yielded classifiers that predicted nearly all older individuals to have PTSD, and those younger to have depression. This 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: $Q(\hat{Y}_D|A = age) \nsim P(Y_D|A = age)$). In the end, the task organizers mitigated the issues by using matched controls – new 50% samples for each class were taken such that the age and gender distributions of both groups matched. Recently, Benton et al. (2017) showed that accounting for demographic factors in the model could leverage this correlation and control for the confounds.

**Assessing Demographic Variance in Language.** A final case study in applying the predictive bias framework demonstrates where inferring user demographics can be helpful in mitigating 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 is expected to be representative (i.e. surveys) while the *features* come with biases – the source feature distributions with respect to human attributes are dissimilar than the ideal distribution while the source outcomes match that target outcomes ($Q(X_{source}|A) \nsim P(X_{target}|A)$ but $Q(Y_{source}|A) \sim P(Y_{target}|A)$). This means that, for this case, the effectiveness of countermeasures to prevent selection and semantic biases to features ($X_{source}$ and $X_{target}$) should result in increased predictive performance against a representative community outcomes. Indeed, Giorgi et al. (2019) adjusted estimates of features, $X$, to match representative demographics and socio-economics by using inferred user attributes, and found improved predictions of the life satisfaction of a community from Twitter.

## 7 Conclusion

We have presented a comprehensive overview of the recent literature on bias in NLP and related fields. Based on this survey, we have developed a unifying conceptual framework to describe bias sources and effects (rather than just the effects), which allows us to group and compare similar works on countermeasures.

While the scope of this paper can give the impression that bias is a growing problem, we would like to point out that bias is not necessarily something that has gone awry, but rather something nearly inevitable in statistical models. We do, however, need to acknowledge and address it with proactive measures. Having a formal framework of the causes can help us achieve this.

As main points, we would like to leave the reader with these: (1) every predictive model with error is bound to have disparities over human factors (even those not directly integrating human factors); (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 factors" (or human factors 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 and hope it inspires further work in both identifying and countering bias, as well as conceptually and mathematically defining bias in NLP.

## 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*.

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.

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.

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, 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 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, 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.

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.

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.

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.

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 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.

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

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.

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.