Bias Identification and Attribution in NLP Models With Regression and Effect Sizes

Erenay Dayanik, IMS, University of Stuttgart, Germany erenay.dayanik@ims.uni-stuttgart.de
Ngoc Thang Vu, IMS, University of Stuttgart, Germany ngoc-thang.vu@ims.uni-stuttgart.de
Sebastian Padó, IMS, University of Stuttgart, Germany sebastian.pado@ims.uni-stuttgart.de

Abstract There is a growing awareness that many NLP systems incorporate biases of various types (e.g., regarding gender or race) which can cause significant social harm. At the same time, the techniques often used for the statistical analysis of biases in NLP systems are still relatively basic. Typically, studies test for the presence of a significant difference between two levels of a single bias variable (e.g., gender: male vs. female) without attention to potential confounders, and do not quantify the importance of the bias variable. This article proposes to analyze bias in the output of NLP systems using multivariate regression models. Such models provide a robust and more informative alternative which (a) generalizes to multiple bias variables, (b) can take covariates into account, (c) can be combined with measures of effect size to quantify the size of bias. Jointly, these effects contribute to a statistically more robust identification and attribution of bias that can be used to diagnose system behavior and extract informative examples. We demonstrate the benefits of our method by analyzing a range of current NLP models on two tasks, namely one regression task (emotion intensity prediction) and one classification task (coreference resolution).

1 Introduction

Machine learning has been a major driver of innovation in natural language processing since the 1990s, but only the last decade has seen the widespread deployment of NLP methods for use by non-experts: Applications such as neural machine translation (Wu et al., 2016) or voice assistants (Këpuska and Bohouta, 2018) are now routinely available through end users' mobile phones, and NLP methods are increasingly used in domains outside computer science such as police work (Sun et al., 2021) and recruiting (Singh et al., 2010).

Such systems are, from a user perspective, black boxes whose predictions are generally taken at face value. This makes the question pertinent to what extent the machine learning methods underlying these NLP models are *fair*, or, on the contrary, to what extent they are subject to *biases* which impact their predictions. More formally, Friedman and Nissenbaum (1996) defined biased computer systems as systems that "*systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others*"; (see Mehrabi et al. (2021) for a very similar definition). Clearly, such biases have the potential to cause concrete *harm* for the disadvantaged groups or individuals (Bender and Friedman, 2018; Blodgett et al., 2020) and must be observed and controlled as far as possible.

A practical aspect of bias analysis, which the above definition leaves open, is whether discrimination is measured "in vitro" (at the level of system performance) or "in vivo" (at the level of real world consequences). In line with the majority of NLP studies on bias, the present study focusses on bias measured "in vitro", i.e., in the form of systematic differences in system performance across groups. We acknowledge the need to better understand how such "in vitro" bias translates into "in vivo" real-world consequences, and argue below that the methods we propose offer a first step in this direction.

A quickly growing body of studies has indeed found that biases are, unfortunately, pervasive in NLP systems (Mehrabi et al., 2021). One of the first studies on bias, Bolukbasi et al. (2016) analyzed similarity relations in word embeddings and found a substantial *gender bias*, as a result of which, e.g., *woman* was more similar to *nurse* than *doctor*, while *man* was more similar to *doctor* than *nurse.* Davidson et al. (2019) found systematic and substantial racial biases in five Twitter datasets annotated for offensive language detection, where African American English tweets were overclassified as hateful compared with Standard American English, and Díaz et al. (2018) found a significant age bias in many sentiment analysis algorithms, attributing less positive attitudes to older participants. See Section 2 for more details.

Consequently, dealing with biases is rapidly becoming a major high-level consideration in the design and development of NLP systems. The three main bias-related tasks are (a) bias identification (is bias present?), (b) bias attribution (where does the bias come from?) and (c) bias mitigation (how to minimize the bias?). In this article, we focus on the first two tasks, bias identification and attribution.

Following the definition given above, the identification of "in vitro" bias involves the establishment of systematic differences in system performance between two parallel stimuli sets for different levels of a *bias variable* such as gender or race. Put simply, the question is: Does, e.g., the gender of an author have a systematic influence on the output of an NLP system (e.g., are texts written by women predicted to be less positive?), or on the quality of the NLP system? (E.g., are text written by women analyzed less reliably?)

This question can be answered using statistical analysis techniques of increasing complexity, shown in Table 1. To our knowledge, all existing studies on bias fall into either the first or the second group. Studies in the first group only quantify the *performance differences*. For instance, studies investigating gender bias have generated predictions for sentence pairs which differ only in gendered expressions (e.g., cf. Table 2) and reported the difference between these sets (Zhao et al., 2018; Stanovsky et al., 2019). Without considering between-system and between-item variance, it is not clear that such differences are indeed *systematic*, as required by the definition of bias from above. For this reason, studies from the second group additionally carry out *hypothesis tests*, typically t-tests, to assess the statistical significance of the differences (Kiritchenko and Mohammad, 2018).

Although this procedure is conceptually simple and straightforward, it is problematic for two reasons. First, the pairwise hypothesis tests that are being employed in existing work assume that differences between the two sets of stimuli are due to the selected bias variable. They cannot ensure that the putative effect of bias is not due to a *covariate* that acts as a *confounding variable* (McNamee, 2005). For instance, studies on gender bias often use sets of male and female names as part of their stimulus sets (cf. Table 2). Across genders, these names may differ in the average age of the bearer, or simply in their frequency in texts, both of which may influence the performance of NLP systems (Díaz et al., 2018; Gerz et al., 2018). Similarly, author gender may be correlated with topic (Schmid, 2002; Schwemmer and Jungkunz, 2019), which can also have an impact on analyses. Therefore, even when an analysis of performance differences by gender may yield a significant performance difference, it is advisable to rule out that there are competing explanations of the difference in performance in terms of other factors.

Second, bias studies in NLP currently generally test for *statistical significance*, but very few consider *model fit* and *effect sizes* (with the notable exception of (Caliskan et al., 2017)). Significance ensures that an identified effect is not a random fluke, but does not quantify how much of the variance in the predictions is due to the bias. Given a sufficiently large dataset, even very small differences that are not practically relevant can reach significance. In contrast, the computation of effect sizes permits users to understand the practical impact of biases (Sullivan and Feinn, 2012), and is therefore arguably a first step moving from bias "in vitro" towards bias "in vivo".

In this article, we propose that these two limitations can be alleviated by adopting *multivariate regression models* such as linear and logistic regression for bias identification. This solution has already become standard in neighboring disciplines like linguistics and psychology. In regression models, bias variables and their covariates form the independent variables, and the predictions of NLP systems for corresponding instances constitute the dependent variable of the equation. As the last column in Table 1 presents, multivariate regression models have many advantages over the other two approaches for bias analysis: (a), they generalize to multiple bias variables; (b), they offer a principled treatment of covariates; (c), they come with measures of effect size that quantify the size of the bias, and (d), they provide a rich diagnosis of system behavior and can be mined easily to extract informative datapoints. In NLP, regression models of various kinds have been used widely as *predictive* models. In our paper, we focus on their use as *explanatory* models, where the focus is on building an interpretable model. Models of this type have been applied to analyze the influence of task and data properties on the performance of sequence labeling models (Papay et al., 2020) or the influence of various textual properties of author responses on the peer review process (Gao et al., 2019). We would like to stress that the goal of this procedure is not to "explain away" biases, but rather to propose a more stringent procedure to identify them, in order to strengthen their empirical standing.

Our concrete contributions are as follows:

- We identify limitations of the statistical methods that are currently applied for bias identification (Section 1).
- We propose a workflow and a set of best practices for designing, computing and interpreting multivariate regression models for this task (Section 3).
- We apply our workflow to two tasks: emotion intensity prediction, a regression task (Section 4) and coreference resolution, a classification task

	Performance Difference (Rudinger et al. 2018, Zhao et al. 2018, etc.)	Performance Difference plus Hypothesis Testing (Caliskan et al. 2017, Kiritchenko et al. 2018, etc.)	Regression Modeling with Effect Sizes (Ours)
Assessing statistical significance	-	+	+
Quantifying the impact of multiple variables	-	-	+
Diagnosing system behavior	+	+	+

Table 1: Comparison of different approaches to statistical analysis of bias.

(Section 5). Our results are in line with established findings, but permit a more nuanced and richer understanding of system behavior.

The complete code for our experiments is publicly available at `https://github.com/multireg/multireg-effect`.

2 Related Work

This section sketches the state of the art in bias analysis, More comprehensive reviews are provided by Sun et al. (2019), Blodgett et al. (2020) and Mehrabi et al. (2021).

Bias in embeddings. At the representation level, almost all state-of-the-art NLP systems use corpus-derived embeddings. These embeddings were the starting point for a lot of work on bias in NLP. Bias in embeddings is generally shown by comparing embeddings for two sets of previously established, e.g., gendered (male and female) words (e.g. *man, woman*). Bolukbasi et al. (2016) define the gender bias of a word by its projection on the difference vector between male and female embeddings; this method was found by Gonen and Goldberg (2019) to be an imperfect metric of bias. As an alternative, the WEAT benchmark (Caliskan et al., 2017) defines bias in terms of similarity to the two sets of gendered words and uses a statistical hypothesis test to assess the statistical significance of the difference. Later, WEAT was used for measuring other bias types (e.g. Race) as well. Caliskan et al. (2017) in fact use effect sizes as a metric, but this was not taken up by follow-up work in NLP such as Gonen and Goldberg (2019).

Going beyond gender, Garg et al. (2018) analyzed ethnic biases in historical embeddings covering 100 years of language use. Swinger et al. (2019) showed that word embeddings of names reflect broad societal biases that are associated with those names, including race, gender, and age biases. Comparable biases also have been demonstrated in multilingual embeddings (Lauscher and Glavaš, 2019; Zhao et al., 2020). The perspective on types and sources of bias is continuing to broaden; Hovy and Prabhumoye (2021) propose a taxonomy of five sources of bias in NLP systems, namely the data, the annotation process, the input representations, the models, and the research design.

Bias in NLP systems. At the system level, bias has been investigated in applications including named entity recognition (NER), Machine Translation (MT), Sentiment Analysis, and Coreference Resolution. Kiritchenko and Mohammad (2018) examined 219 sentiment analysis systems and found that a majority exhibits gender and race biases. Mehrabi et al. (2019) reported that NER models recognize male names with higher recall compared to female names. Rudinger et al. (2018) and Zhao et al. (2018) showed that coreference resolution systems perform unequally across gender groups by associating occupations (such as doctor and engineer) more with men and others (like nurse) more with women. Similarly, Stanovsky et al. (2019) found that both commercial and academic MT models are at risk of generating translations based on gender stereotypes rather than the actual source content.

Bias in systems is usually measured by using benchmarks datasets for specific tasks with a one-factor design which are created to be as balanced as possible while varying the levels of the bias variable. Examples include WinoBias (Zhao et al., 2018) and WinoGender (Rudinger et al., 2018), two benchmarks for gender bias in coreference resolution which contrast "pro-stereotype" cases (the correct antecedent of a pronoun is conventionally associated with the pronoun's gender) and "anti-stereotype" cases (opposite situation); GAP (Webster et al., 2018), a dataset for the same task described in detail in Section 5; and the Equity Evaluation Corpus (EEC, Kiritchenko and Mohammad (2018)), developed to analyze gender and race bias in sentiment analysis and described in detail in Section 4. Bias is then quantified by measuring the differences in performance between these levels. Sometimes, but not always, the differences are subsequently tested for statistical significance, e.g. t-tests. To our knowledge, almost no studies on system-level bias have considered covariates, nor computed effect sizes, which makes them vulnerable to the criticisms outlined in Section 1.

An exception is a recent study Feder et al. (2021) which, like ours, disentangles bias from confounding factors. However, instead of performing correlational analysis of model predictions, they aim at full-fledged causal analysis. Since causal relations can often not be recovered from data (Pearl, 2009), they assume that a causal graph modeling dependencies between predictors are given by a domain expert and show how to fine-tune contextualized embedding models with adversarial training to minimize bias. Thus, the two studies take complementary approaches: Feder et al. (2021) applies to model construction, while our study carries out black-box analysis of existing models.

Bias Mitigation. There are two main families of methods to mitigate bias at the representation level. Approaches from the first family create a modified version of the original data set that is biased in the opposite direction, training models on the union (Park et al., 2018; Zhao et al., 2019; Stanovsky et al., 2019). Approaches from the second family mitigate bias by transforming learned embeddings according to some balancing objective (Lauscher and Glavaš, 2019; Kaneko and Bollegala, 2019; Dev et al., 2020; Kaneko and Bollegala, 2021a,b).

At the system level, Zhao et al. (2017) proposed to constrain model predictions to follow a distribution from a training corpus. Rather than constraining the output, some of the previous work such as Elazar and Goldberg (2018); Zhang et al. (2018) and Kumar et al. (2019) used adversarial learning to remove unintended bias from the latent space during model training. Adjusting the loss function is another popular system level approach for bias mitigation. For instance, Qian et al. (2019) introduces a new term to the loss function to equalize the probabilities of male and female words in the output, and Jin et al. (2021) introduce a regularization term which reduces the importance placed on surface patterns.

Note that almost all mitigation methods require knowledge about which variables are (potentially) introducing bias, underlining the importance of reliable identification of bias variables.

3 Bias Identification With Regression Models: A Workflow

Following the discussion in the previous sections, the task of ("in vitro") bias identification is to establish that a bias variable – in contrast to other covariates which act as confounders – is primarily responsible for systematic variance in an observed variable, namely the performance of some computer system.

This is, of course, a very general task that arises in many empirical fields. A prominent family of techniques to address this task is *matching* (Rubin, 1973), which aims at generating two datasets that differ in the bias variable, but are as close as possible in their distribution over the covariates, so that any difference between the two datasets can be attributed to the bias variable. Matching is widely used in social sciences, economy, and medicine and many specific methods exist; see Stuart (2010) for an overview.[1]

Importantly, matching takes place *a priori*, before the experiment is carried out. This poses two challenges for applications in natural language processing: (a), dataset creation is dependent on the selection of covariates, so that it is not possible to assess the impact of new covariates on existing datasets without loss of comparability; (b), matching samples from the set of all datapoints, creating controlled rather than natural datasets, which may conflict with the desideratum of estimating model performance in broad-coverage scenarios.

The alternative is to carry out a *post-hoc* analysis that assesses the effects of the various covariates. The intuition is to start from a simple pairwise comparison of two levels of a bias variable (cf. the first and second column in Table 1) and add covariates to see whether the effect of the bias variable remains unaffected. This procedure has become standard in the last decade in neighboring fields like linguistics and psychology which have moved from significance tests (Student's t-test, analysis of variance) to the family of *multivariate regression models* (Bresnan et al., 2007; Baayen, 2008; Jaeger, 2008; Snijders and Bosker, 2012). Regression models estimate the relationships between the dependent (previously called observed) variable – in this case, system performance – and one or more independent variables – in this case, the putative bias variable and its covariates, each of which is assigned a direction and a significance. Since dataset creation is dependent from covariate analysis, regression models can be used to test new candidates for confounders on existing datasets.

At this point, it can be whether the fundamentally linear regression models are the right tool for the job, in particular given the broad success of non-linear deep learning models in NLP over the last years. We believe that it makes sense to distinguish carefully between the task of *output prediction* (given language input, predict language output) on which non-linear models indeed excel and the task of *performance prediction* (given [meta data for an] input and a model, predict how well the model does on the input). The latter is a considerably simpler problem which permits the use of linear models, as evidenced by a number of successful studies taking this approach (Beinborn et al., 2014; Papay et al., 2020; Caucheteux and King, 2022).

This section provides a practical workflow to set up a regression model for bias analysis, shown in Figure 1. Our starting point is the presence of a dataset with system predictions. Step 1 is the selection of an appropriate regression model. In Step 2, we choose a set

[1] Note that the term *bias* is used differently in the matching literature, namely as the effect of confounders on the observed variable.

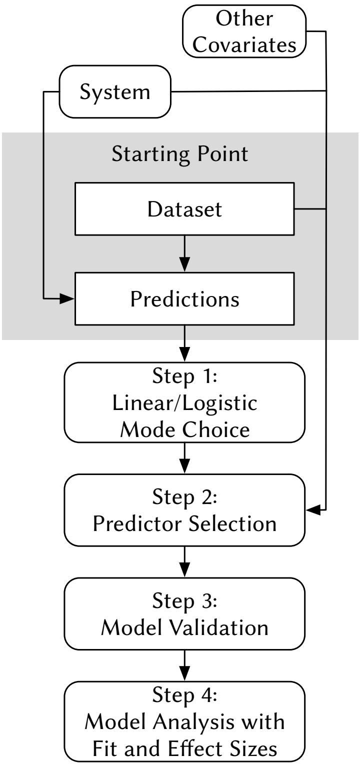

Other Covariates
System
Starting Point
Dataset
Predictions
Step 1: Linear/Logistic Mode Choice
Step 2: Predictor Selection
Step 3: Model Validation
Step 4: Model Analysis with Fit and Effect Sizes

Figure 1: Workflow for regression-based bias analysis

of predictors with the potential to systematically influence the predictions of the systems, (i.e., the putative bias variable and plausible confounders) and carry out a regression analysis. Next, Step 3, model validation, ensures that the regression model is well specified and interpretable. Finally, Step 4 utilizes effect size analysis methods to explore how much of the system predictions can be attributed to the influence of the predictors.

Running example. We will illustrate the steps of the workflow on an actual (non-NLP) example, namely the effect of smoking on mortality, a topic of long-running interest in public health that has been analyzed extensively with regression models. The most basic finding is that smoking, overall, causes a strong increase in mortality (Doll et al., 2004). Why it is still reasonable to carry out a regression analysis in this case is that other lifestyle choices (alcohol consumption, diet, etc.) also presumably influence mortality, but exhibit correlations (Padrão et al., 2007). These are sometimes surprising – e.g., Tjønneland et al. (1999) found a correlation between wine and healthy diet. At the same time, approaches like matching are not applicable since the lifestyle properties of the participants cannot be influenced retroactively.

3.1 Step 1: Choice of Regression Model

The most common two forms of regression analysis are linear regression and logistic regression. When used to analyze the output of computational models, linear regression is appropriate to analyze the output of regression tasks, and logistic regression for the output of classification tasks.

Linear regression predicts the outcome of a continuous random variable y as a linear combination of weighted predictors x_i:

$$y \sim \alpha_1 x_1 + \cdots + \alpha_n x_n \tag{1}$$

where the coefficients α_i can be interpreted as the change in y resulting from a change in predictor x_i, keeping the other predictors constant.[2]

In contrast to linear regression, logistic regression does not model the outcome of the binary random variable y directly. Instead, it models the probability $P(y = 1)$, assuming that $P(y = 1)$ stands in a linear relationship to the logistically transformed linear combination of weighted predictors:

$$P(y = 1) \sim \sigma(\alpha_1 x_1 + \cdots + \alpha_n x_n) \tag{2}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function. Here, the coefficients α can be interpreted as the change in the logit for a unit change in the predictor.

Both types of regression support continuous, binary, and categorical predictors; the latter type is generally represented as a set of binary indicator predictors. As indicated above, these models assume that the predictors have an additive effect on the dependent variable (in the linear case) or its logit (in the logarithmic case).

Running example. In our mortality example, the outcome of the regression model is (some variant of) a death rate. Depending on the exact choice of measure, it might be appropriate to choose a linear regression model, when the death rates are approximately normally distributed (Gardner, 1973); or it might be appropriate to choose a logistic regression model, when the death rates can be interpreted as probabilities (Zhu et al., 2015b).

3.2 Step 2: Selection of Predictors

Maybe the most central step in the use of a regression model for bias analysis is the selection of the set of predictors for the regression model – that is, the putative bias variable and a set of plausible confounders to assess the respective roles of these variables in explaining the variance of the dependent variable.

[2] If the dependent variable is not (approximately) normally distributed, other types such as Poisson or negative binomial regression may be more appropriate.

This task is the responsibility of the user and typically involves domain knowledge. Typically, a user carrying out a bias identification analysis will have one (or a small number) of bias variables in mind, but need to select plausible confounders.

The five primary sources of bias variables given by Hovy and Prabhumoye (2021) can also serve as sources of confounders. The most straightforward of these are *data* and *input representations*, that is, properties of the text underlying the model, many of which are known to impact model performance. For example, low-frequency words and classes are modeled less reliably, longer stretches of text are harder to analyze, and so on (Poliak et al., 2018; Dayanik and Padó, 2020). Similarly, differences among *annotators* (age, social and cultural background, task familiarity) can impact model performance through labeling decisions (Sap et al., 2019), and obviously design decisions of the *system*, such as the choice of neural network architecture, contribute as well (Basta et al., 2019). Hovy and Prabhumoye's fifth category of *research design* is least relevant for our purposes, since it is concerned with systematic gaps in the field as such rather than analysis of individual studies.

Thus, for many problems, there will be a range of theoretically motivated covariates. The actual analysis will proceed in an interlocking fashion between exploratory data analysis based on domain knowledge – to identify interesting candidates for covariates – and regression modeling – to obtain statistically sound assessments of these covariates. In practical terms, the limiting factor is often that covariates need to be available as annotation on the dataset under consideration. While this is often relatively simple for the domains of input representation and systems, and doable for the domain of data, only recently has natural language processing started to record and analyze annotator properties (Sap et al., 2019), and there is an inherent tension between insights into annotation biases and annotator privacy. In some cases, however, covariates can be obtained by automatic or semi-automatic means. As an example, see our estimation of the typical age for the bearer of a specific first name on the basis of census data in Experiment 1 below. Such approaches can ease the burden of data collection, but the analysis should take into account the uncertainty introduced by automatic annotation.

Running example. In our lifestyle example, the covariates ideally include as many lifestyle factors as possible (such as alcohol consumption, diet, exercise, occupational hazards) as well as environmental factors (housing, climate) and personal factors such as family history of certain diseases. In practice, again, only a limited range of such factors is likely to be available.

3.3 Step 3: Model Validation

While regression models technically support arbitrary covariates, strong correlations among predictors, so-called *multicollinearity*, can distort the estimation of coefficients to the point that predictors are suggested to be significant when they are not, and vice versa (McNamee, 2005). Therefore, models should be checked for the presence of multicollinearity. There is a wide range of tests available, see Imdad Ullah et al. (2019) for a recent overview. We use the so-called variance inflation factor (VIF). VIF measures how much the variance of a predictor's coefficient is inflated due to correlations with other predictors. The VIF is computed for each independent variable V_i as

$$\mathrm{VIF_i} = 1/\big(1 - \mathrm{R_i^2}\big) \tag{3}$$

where R_i^2 is the correlation coefficient obtained when predicting α_i from all other predictors. Thus, the more collinearity is present, the higher VIF_i. VIF values of 4 or greater indicate severe multicollinearity, and values above 2.5 call for further investigation (Salmerón et al., 2018). In this case, a number of strategies are available, including dropping covariates, dimensionality reduction, and regularization methods (see Dormann et al. (2013) for details).

Another possible component of model validation is *predictor (feature) selection* based on an analysis of feature contributions. In many NLP tasks, irrelevant or unimportant features are removed for reasons of efficiency or to avoid overfitting (Li et al., 2009). In fields like psychology, where models serve explanatory purposes, predictor selection is discussed more controversially (Barr et al., 2013; Bates et al., 2018). In bias analysis, the goal is to test whether the effect of the putative bias variable stands up to the addition of covariates – the more covariates added to the model while retaining a significant contribution of the bias variable, the stronger the evidence for a specific role of the bias variable. For this reason, we believe that regression based bias analysis should be carried out on a comprehensive set of predictors, without feature selection (Barr et al., 2013).

Running example. In our lifestyle example, is it arguably important to check for multicollinearity, since the various covariates may be predictive of one another. For example, cramped housing conditions and occupational hazards are strongly linked through the shared cause of poverty (Hajat et al., 2015).

3.4 Step 4: Computing Model Fit and Effect Sizes

The coefficients α computed by regression models (cf. Step 1) are accompanied by indications of the confidence level at which they are different from zero (i.e., whether the predictor has a significant effect). Furthermore, the global quality of regression models can be assessed by a number of statistics. Among them, we use *goodness of fit* which describes the proportion of the variance in the data that is explained by the independent variables of a regression model. The goodness of fit of a linear regression model is measured by R^2:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{4}$$

where $\hat{y}_i$ is the model's prediction for data point i and $\bar{y}$ is the mean of the observations.

In logistic regression, there is no exact equivalent of R^2. Among several pseudo R^2 measures that have been proposed, Aldrich-Nelson pseudo-R^2 with Veall-Zimmermann correction (R^2_{VZ}) most closely approximates the R^2 in linear regression (Smith and Mckenna, 2013):

$$R^2_{\mathrm{VZ}} = \frac{2[\mathrm{LL(Null)} - \mathrm{LL(Full)}]}{2[\mathrm{LL(Null)} - \mathrm{LL(Full)}] + N} \frac{2\mathrm{LL(Null)} - N}{2\mathrm{LL(Full)}} \tag{5}$$

where LL(Full) and LL(Null) are the log-likelihood values for the model with all predictors and for the empty model (without predictors), respectively.

Goodness of fit measures the overall ability of the model to explain the dependent variable. *Relative importance*, on the other hand, refers to the contribution of individual predictors (Achen, 1982). While assessment of relative importance in linear models with uncorrelated independent variables is simple (the impact of each predictor is its R^2 in univariate regression), in real-world datasets variables are generally correlated, as a result of which their impacts are not additive (Grömping, 2006). Lindeman-Merenda-Gold (LMG) scores (Lindeman et al., 1980) and Dominance Analysis (Budescu, 1993) are two popular techniques to figure out the individual contributions to the R^2 of the model of the predictors in linear and logistic regression, respectively.

The LMG method adds predictors to the regression model sequentially, and considers the resulting increase in R^2 as its contribution. Since this method depends on the possible orders in which predictors are added, the LMG score of a predictor x_k when added to a model with a set of predictors P is defined as the average of the increase in R^2 when adding x_k to all subsets of P (Grömping, 2006):

$$\mathrm{seq}\, R^2(M|S) = R^2(M \cup S) - R^2(S) \tag{6}$$

$$\mathrm{LMG}\,(x_k) = \frac{1}{n} \sum_{j=0}^{p-1} \sum_{\substack{S \subseteq P \\ n(S)=j}} \frac{\mathrm{seq}\, R^2(\{x_k\}|S)}{\binom{p-1}{j}} \tag{7}$$

where $R^2(S)$ corresponds to the goodness of fit measure of a model with regressors in set S (cf. Eq 1) and $\mathrm{seq}\, R^2(M|S)$ refers to the increase in R^2 when the regressors from M are added to the model based on the regressors S.

For logistic regression, there is again no direct counterpart. We propose Dominance Analysis (Budescu, 1993) as a measure of the relative importance of each predictor. Dominance analysis considers one predictor (x_i) to completely dominate another (x_j) if x_i's additional contribution to every possible model which does not include these two predictors is greater than contribution of x_j. In cases where complete dominance cannot be established, general dominance can also be used. One predictor generally dominates another if its average conditional contribution over all model sizes is greater than that of the other predictors (Azen and Traxel, 2009).

We propose the following interpretations for the regression scores outlined above: (a) At the system level, R^2 and pseudo-R^2 are indicators of the amount of variance in the system predictions that can be explained by the predictors and measure the *systematic bias* of a system. (b) At the predictor level, the significance of a predictor indicates the *presence of a specific bias*, and its effect size measures its *practical impact*; (c) the sign of a coefficient indicates the *direction* of a bias.

Regarding (b), an important difference between the application of significance testing in bias analysis and the usual use in NLP to compare competing models is that in our case, null results are arguably informative: they indicate the *absence* of a particular bias, according to the standards of significance. Naturally, the usual disclaimers regarding null results apply: care should be taken to ensure that they are not the result of faults in the experimental setup.

Running example. In our lifestyle example, the outcome of this step is a better understanding of individual risk factors, such as smoking, as opposed to the cluster of 'smoking and associated factors' that is obtained from a simple smoker-vs.-non-smoker analysis. Such an understanding is crucial to better assess the risk of individual patients based on their individual risk profile which might include compounding factors (high blood pressure, alcohol consumption) or mitigating factors (exercise, healthy diet). Again, note that the goal of this analysis is not to detract from the hazardous nature of smoking, but to better estimate of the effects of the relevant predictors on the outcome, namely mortality.

4 Experiment 1: Emotion Intensity Prediction

We now employ regression models to reanalyze model predictions on two experiments on standard datasets from the bias literature using the workflow defined in

Template
1. [PER] feels [EMO].
2. The situation makes [person] feel [EMO].
3. I made [person] feel [EMO].
4. [PER] made me feel [EMO].
5. [PER] found herself in a [EMO] situation.
6. [PER] told us about the recent [EMO] events.
7. The conversation with [person] was [EMO].
8. I saw [person] in the market.
9. I talked to [person] yesterday.
10. [PER] goes to school in our neighborhood.
11. [PER] has two children.

African American		**European American**	
Female	**Male**	**Female**	**Male**
Ebony	Alonzo	Amanda	Adam
Jasmine	Alphonse	Betsy	Alan
Lakisha	Darnell	Courtney	Andrew
Latisha	Jamel	Ellen	Frank
Latoya	Jerome	Heather	Harry
Nichelle	Lamar	Katie	Jack
Shaniqua	Leroy	Kristin	Josh
Shereen	Malik	Melanie	Justin
Tanisha	Terrence	Nancy	Roger
Tia	Torrance	Stephanie	Ryan

Table 2: Sentence templates in EEC dataset (top) and female and male first names associated with being African American and European American (bottom). [EMO]: an emotion adjective

Section 3.

Our first experiment is concerned with emotion intensity prediction. This task aims at combining discrete emotion classes with different levels of activation. Given a tweet and an emotion, the task requires to determine a score between 0 and 1 which is the intensity expressed regarding an emotion. Emotion intensity prediction was among the first NLP tasks to receive attention from a bias angle, when Kiritchenko and Mohammad (2018) found that among more than 200 emotion intensity prediction systems, almost all were biased with regard to gender or race. (In the remainder of the article, we will use 'system' to refer to models performing the task at hand, and 'model' to refer to the regression models we use for analyzing the systems' performance.)

4.1 Dataset and Previous Analysis

We use EEC, the same dataset used for the large-scale bias analysis of sentiment analysis mentioned above (Kiritchenko and Mohammad, 2018). EEC is a bias analysis benchmark created to evaluate fairness in sentiment analysis systems. It consists of 11 sentence templates instantiated into 8,640 English sentences for four emotions (anger, joy, fear, sadness). Instantiated templates differ only in the name.[3] The dataset compares (a) male vs. female first names, and (b) European American vs. African American first names, using ten names of each category. Table 2 shows examples of such template sentences along with names that tend to belong to African American or European American demographic groups.

	train	dev	test	task
EI-reg	1701	388	1002	EIP
EEC	-	-	2100	EIP
GAP	-	2000	2000	CR

Table 3: Number of examples in the datasets used in our emotion intensity prediction (EIP) and coreference resolution (CR) experiments.

Kiritchenko and Mohammad (2018) used the EEC as a secondary test set for systems submitted to the SemEval 2018 Task 1 (Mohammad et al., 2018). For each system, they compared the average emotion intensities across different demographic groups using t-tests. They found that almost all systems consistently scored sentences of one gender and race higher than another, but bias directions were not consistent: e.g., some systems assigned higher emotion intensities to African Americans and lower ones to European Americans, while others show the opposite behavior. This apparently random behavior of the systems has no clear explanation and arguably raises concerns about a possible role of randomness in the analysis.

4.2 Systems

Since the predictions of the systems that participated in SemEval 2018 Task 1 are not publicly available[4], we instead implement and analyze five systems ourselves. Four systems represent the main architectures submitted to the shared task (Kiritchenko and Mohammad, 2018): A SVM unigram baseline and three neural systems based on word2vec word embeddings. To extend the model set to the current state of the art (2021), we include a transformer-based architecture as fifth system.

Support Vector Machine (SVM) We implement the unigram-based SVM used as baseline system in Mohammad et al. (2018).

[3] The EEC templates can also be instantiated using gendered noun phrases, but since these are unspecific with regard to the race variable, we focus on the version with proper nouns. This corresponds to the race analysis of the original study.

[4] Personal communication with the authors of shared task.

Convolutional Neural Network (CNN) Based on Aono and Himeno (2018), this system predicts an intensity score by first performing convolutions of different sizes on input word embeddings, followed by max-pooling and a shallow multi-layer perceptron (MLP).

Recurrent Neural Network (RNN) Our RNN is comparable to Wang and Zhou (2018). A two-layer BiLSTM traverses the input. The final hidden states in both directions from the final layer are concatenated and fed to a fully connected layer.

Attention Network (ATTN) This system is based on a CNN-LSTM architecture with attention similar to Wu et al. (2018). The input is fed to a single-layer BiLSTM. Next, an attention mechanism weights the hidden states, which are then passed through a CNN. The outputs of the CNN feature maps are concatenated and passed through a pooling layer and two fully connected layers.

Transformer-Based Neural Network (BERT) This system is based on the BERT$_{BASE}$ multilayer bidirectional Transformer architecture (Devlin et al., 2019). It adds a linear layer on top of BERT and uses the final hidden state of the special [CLS] token as the latent representation of the input tweet, inspired by May et al. (2019).

We train and evaluate all the systems on the Anger partition of the EI-reg corpus (Mohammad and Bravo-Marquez, 2017) and EEC respectively. EI-reg was created by querying tweets in three languages (English, Arabic, Spanish) and for four emotions (Anger, Fear, Joy, Sadness) with words that were associated with the emotion at different intensity levels, such as *angry, annoyed, irritated* for Anger. Table 3 shows data statistics for both datasets.

4.3 Setup of the Regression Model

Bias Variable. In the EEC setup, the input sentences differ only in the person names that are filled in. We use the same two bias variables considered by the original study, namely Race and Gender.

Covariates. Due to the minimalist nature of the templates, coupled with the fact that the only part of the templates that is manipulated across conditions is the names, there is a limited range of linguistic properties that can systematically covary with bias. We consider two that we consider promising candidates. The first one is the (perceived) Age of a name is computed as the mean age for each name from US Social Security data.[5]

Example	Properties				Intensity
	Gender	Race	Age	Freq	
Frank feels angry	Male	EA.	Old	0.05	0.55
Alonzo feels angry	Male	AA.	Old	0.24	0.48
Justin feels angry	Male	EA.	Yng	0.27	0.46
Lamar feels angry	Male	AA.	Yng	0.42	0.49
Jasmine feels angry	Female	AA.	Yng	0.47	0.47
Ellen feels angry	Female	EA.	Old	0.19	0.50

Table 4: Example sentences for the first template from Table 2 with their properties (EA.: European American, AA.: African American, Yng: Young). Intensity predicted by the the RNN system.

We discretize age, using 40 as the young/old boundary, following the assumption that 'older' names occur in different contexts than 'younger' names. The second covariate is the linguistic frequency of the name in the training data, since low-frequency names have found to be a source of low performance in NLP models (Dayanik and Padó, 2020). Since no explicit frequencies are available for the Google News skipgram vectors (Mikolov et al., 2013), we approximate frequency by vector length, which correlates highly with frequency (Roller and Erk, 2016). This is different from the 'real world' frequency of the name, which arguably is less likely to reflect in the behavior of an NLP model. Table 4 shows examples from the EEC with their properties.[6]

Model Shape We analyze the intensities predicted by our systems as in the original study, performing linear regression analysis at the level of each template with the following model:

$$\text{Intensity} \sim \text{Race} + \text{Gender} + \text{Age} + \text{Freq} \qquad (8)$$

For Race, 1 means African American and 0 European American. For Gender, 1 means male and 0 female. For Age, 1 means young and 0 old.

Recall that on this task, there is no right or wrong answers. Instead, the focus of interest is whether the systems assign different intensities to a template dependent on the properties of the instantiating name. If they do not, none of the predictors will show a significant effect; if they do, significant effects will emerge.

Model Validation. Table 5 shows the variance inflation factors for the variables. Since only a single VIF value is larger than 2.5, and only marginally so, we conclude that multicollinearity is not a problem.

[5] We use data from `https://bit.ly/34cgjki` and the methodology from `https://bit.ly/30f8lps`.

[6] We also performed experiments using a non-discretized version of age and including real-world frequency. We observed a substantially similar outcome (same levels of significance, coefficient signs for predictors, and almost the same overall R^2 values).

	Race	Gender	Frequency	Age
VIF	2.03	1.42	2.68	1.29

Table 5: VIF scores for the full set of variables.

		CNN	RNN	ATTN	BERT	SVM
	Coef.	-0.010^{*}	-0.010^{*}	−0.002	−0.008	0.001
R	Abs. LMG	0.080	0.082	0.010	0.068	0.018
	Per. LMG	0.42	0.47	0.06	0.48	0.03
	Coef.	0.006	0.002	0.001	−0.001	-0.003^{***}
G	Abs. LMG	0.037	0.003	0.020	0.025	0.523
	Per. LMG	0.20	0.02	0.12	0.18	0.86
	Coef.	0.005	0.001	0.001^{*}	−0.003	0.001
A	Abs. LMG	0.049	0.060	0.070	0.027	0.014
	Per. LMG	0.26	0.34	0.40	0.19	0.02
	Coef.	0.016	0.019	0.015	0.010	−0.001
F	Abs. LMG	0.023	0.029	0.073	0.021	0.048
	Per. LMG	0.12	0.17	0.42	0.15	0.08
R^2 model fit		0.19	0.17	0.17	0.14	0.60

Table 6: Regression-based bias analysis on EEC
(R = Race, G = Gender, A = Age, F= Frequency)
(Abs:Absolute, Per. Percentage)

4.4 Results

Table 6 shows the main results. (We omit intercepts in the table). The columns correspond to systems, and the rows describe the effects of bias variables for each system. For each predictor, we show a coefficient, a confidence level,[7] and an LMG effect size score.

Overall results As discussed in Section 3, we treat R^2 as a measure of systematic bias in a system. Inspection of the R^2 scores indicates that there is a certain amount of systematic bias in all systems, but that the three static-embedding neural systems do a very good job (R^2 between 0.17 and 0.19) compared to the SVM (R^2=0.60). BERT, the only neural system using contextualized embeddings, does an even better job and contains the least amount of systematic bias (R^2=0.14).

Comparison among systems None of the neural systems exhibits a significant gender bias, as the LMG scores show. Unlike Gender, the Race variable is responsible for the significant portion of the amount of variance in the system predictions. The CNN and the RNN systems both show a significant race bias which accounts of about 42–47% (LMG score: ∼ 0.08) of the variance in the intensity predictions. Note that Age, even though it misses significance, also accounts for 25–35% of the variation in intensity in the CNN and RNN. Interestingly, the ATTN architecture shows a different picture: there is a considerable amount of Age bias (40% of variance), but a much smaller race bias; instead, this system shows a frequency bias, which accounts for another 40% of the variance. In the BERT system, none of the bias variable achieve significance. In terms of relative contribution of individual predictors, BERT is more similar to CNN and RNN than to ATTN: Race is still making the largest contribution to the overall bias of the system, with 48%. The SVM differs strikingly: there are hardly any Race and Age biases, but an extremely strong effect of gender (86% of variance). Since this system does not use embeddings, the most likely source of this bias is the training corpus (EI-Reg), as also pointed out by the authors of the original study (Kiritchenko and Mohammad, 2018).

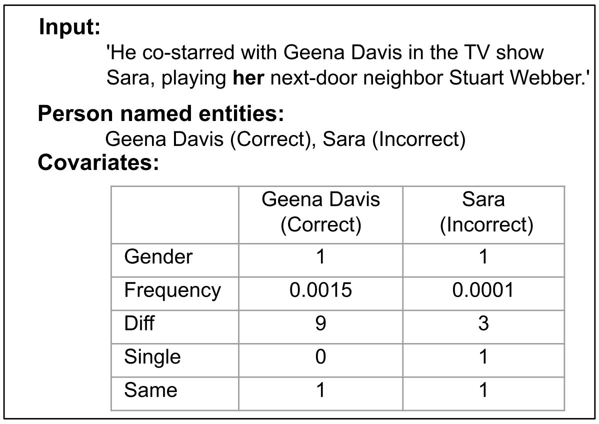
Interpretation While we can confirm the overall race bias found by Kiritchenko and Mohammad (2018), our picture differs substantially: (a) the direction of the bias is consistent among systems: all neural systems predict lower intensity scores for African Americans; (b) we do not observe a significant gender bias among neural systems; (c) we achieve a richer understanding of the systems' predictions, by quantifying the role of these factors, and by adding age and frequency into the picture.

Inspection of Examples Following up on (c), Table 4 presents three pairs of examples from the EEC dataset with their associated intensity values, as predicted by the RNN system. We have selected these instances to highlight the usefulness of the regression model to identify interesting instances. They show that the effect of Race variable (African Americans are assigned lower intensities) can be nullified by age (third example) and frequency (first and second examples). Such considerations remain hidden in an analysis that simply compares means between different groups of predictions.

5 Experiment 2: Coreference Resolution

Our second experiment analyzes several coreference resolvers in order to show how the logistic regression version of our approach can perform bias analysis on classification models. We choose coreference resolution as our task because of its established status in bias analysis; previous work has established that bias, in particular gender bias, is present in numerous coreference systems (Webster et al., 2018; Rudinger et al., 2018; Zhao et al., 2018). At the same time, coreference resolution, as a discourse level task, is faced with more complex data

[7] We use * for α=0.05, ** for α=0.01, and *** for α=0.001.



Input:

'He co-starred with Geena Davis in the TV show Sara, playing **her** next-door neighbor Stuart Webber.'

Person named entities:

Geena Davis (Correct), Sara (Incorrect)

Covariates:

	Geena Davis (Correct)	Sara (Incorrect)
Gender	1	1
Frequency	0.0015	0.0001
Diff	9	3
Single	0	1
Same	1	1

Figure 2: Example from the GAP dataset.

than more local (i.e., sentence-level) tasks, with a correspondingly larger set of potential confounders. We re-analyze a well-known coreference resolution dataset to verify the presence of gender bias in a manner that is robust against possible covariates.

5.1 Dataset and Previous Analysis

We use GAP (Webster et al., 2018), a human-labeled corpus of ambiguous pronoun-name pairs from English Wikipedia snippets. Each instance in the corpus contains two person named entities of the same gender and an ambiguous pronoun that may refer to either, or neither. System clusters were scored against GAP examples according to whether the cluster containing the target pronoun also contained the correct name (True Positive) or the incorrect name (False Positive). Figure 2 shows an example from the GAP development set (more statistics in Table 3).

In line with previous work (Webster et al., 2018), we use the development set of GAP to carry out our analyses. Below, we report overall system performance on the complete development set, in line with previous work. However, we exclude ≈200 instances from the development set, for which the pronoun does not refer to either of the two candidate named entities, from the regression analysis, since this makes it impossible to compute some of our covariates (cf. Section 5.3).

5.2 Systems

We experiment with six diverse coreference resolvers and analyze their predictions with our approach. As trained versions of all systems were publicly available, we did not need to train any systems ourselves. All systems except the BERT-based one were trained on the English portion of the 2012 CoNLL Shared Task dataset (Pradhan et al., 2012). It contains 2802 training, 343 development documents, and 348 test documents. BERT$_{large}$ Joshi et al. (2020) was pretrained on BooksCorpus (Zhu et al., 2015a) and English Wikipedia using cased *Wordpieces* tokens (Schuster and Nakajima, 2012) and fine-tuned on the 2012 CoNLL ST dataset.

Lee et al. (2013) This system is a collection of deterministic coreference resolution modules that incorporate lexical, syntactic, semantic, and discourse information, incorporating global document-level information. The system won the CoNLL 2011 shared task.

Clark and Manning (2015) This system uses a feature-rich machine learning approach. It performs entity clustering using the scores produced by two logistic classifier-based mention pair classifiers features. Both mention pair classifiers use a variety of common features such as syntactic, semantic and lexical features for mention pair classification.

Wiseman et al. (2016) This was the first neural coreference resolution system which showed that the task could benefit from modeling global features about entity clusters. It uses a neural mention ranker which is augmented by entity-level information produced by a RNN running over the cluster of candidate antecedents.

Lee et al. (2017) This was the first neural end-to-end coreference resolution system that works without a syntactic parser or hand engineered mention detector. It uses a combination of Glove and character level embeddings learnt by a CNN to represent the words of annotated documents. Next, the vectorized sentences of the document are fed into a BiLSTM to encode sentences and obtain span representations. The system also uses an attention mechanism to identify the head words in the span representations. Finally, the scoring functions are implemented via two feed-forward layers.

Lee et al. (2018) This system is an extension of Lee et al. (2017), which improves on two aspects. First, it uses gated attention mechanism which allows refinements in span representations; second, the system applies antecedent pruning which alleviates the complexity of running on long documents. It formed the state of the art for two years.

Joshi et al. (2020) SpanBERT is a variant of the BERT transformer (Devlin et al., 2019) designed to better represent spans of text. It works by (1) masking contiguous random spans, rather than random tokens, and (2) introducing a new objective function called span-boundary objective (SBO) which forces the model to learn to predict the entire masked span from the observed tokens at its boundary. BERT$_{large}$ trained with the SpanBERT

	Gender	C_Freq	C_Diff	C_Single	C_Same
VIF	1.03	1.03	1.88	1.02	1.53
	Gender	I_Freq	I_Diff	I_Single	I_Same
VIF	1.03	1.04	1.58	1.04	1.24

Table 7: VIF scores for the predictors. C_: Correct, I_: Incorrect

method improves the state of the art on many tasks including coreference resolution.

5.3 Setup of the Regression Model

Bias Variable. As in the original study, we use *Gender* as designated bias variable.

Covariates. In contrast to the first experiment, we do not use Age and Race, since the GAP dataset contains numerous named entities that are either not generally known or fictional (such as "the Hulk"). Therefore, these variables are either inapplicable or unknown to the typical annotator. Instead, use discourse-related properties of the antecedents as covariates, since in the task of coreference resolution the structural properties of the discourse arguably play a role in the difficulty of the task:

- *Diff* is the number of tokens between the named entity and target pronoun, normalized by the maximal distance in the corpus;
- *Single* states whether the named entity is a single word or an MWE;
- *Same* indicates whether the pronoun and named entity are in the same sentence;
- *Freq* defines the log-transformed corpus frequency of the entity, computed on the English Wikipedia (*en-wikipedia*) released on 20th March 2019, normalized by the maximal frequency in the corpus. The frequencies for MWEs are calculated based on the syntactic head of the expression.

Since the correct and the incorrect antecedent can differ regarding these properties, each property exists twice. We use the prefix C_for the correct and I_for the incorrect one. For gender, both antecedents have the same gender by design. The bottom part of Figure 2 shows how these covariates are initialized for the given example.

	Male	Female	All	Bias
Lee et al. (2013)	55.4	45.5	50.5	0.82
Clark & Manning (2015)	58.5	51.3	55.0	0.88
Wiseman et al. (2016)	68.4	59.9	64.2	0.88
Lee et al. (2017)	67.2	62.2	64.7	0.92
Lee et al. (2018)	75.9	72.1	74.0	0.95
Joshi et al. (2020)	89.9	87.8	88.8	0.98

Table 8: F_1-Scores of resolvers on the GAP development set (Bias=F_1 Female / F_1 Male)

Model Shape We analyze the performance of the coreference resolvers at the level of individual predictions using following logistic regression model:

$$\begin{aligned} \mathrm{p(Correct)} \sim \sigma(&\mathrm{Gender} + \\ &\mathrm{C\_Freq} + \mathrm{I\_Freq} + \\ &\mathrm{C\_Diff} + \mathrm{I\_Diff} + \\ &\mathrm{C\_Single} + \mathrm{I\_Single} + \\ &\mathrm{C\_Same} + \mathrm{I\_Same}) \end{aligned} \tag{9}$$

where σ is the logistic function. $\mathrm{p(Correct)}$: is 1 if the resolver matches the pronoun with the correct named entity in corresponding instance and 0 otherwise. For Gender, 1 means female and 0 male. For Single, 1 means the entity is a single word, 0 otherwise. For Same, 1 means the entity is in the same sentence as the pronoun, 0 otherwise. We use Dominance Analysis to determine relative importance of each predictor.

In this setup, the regression model predicts whether each of the system predictions is correct or incorrect. To the extent the correctness is affected by the properties of the discourse captured by our predictors, we will obtain significant effects; conversely, should the correctness be fully random or dependent on properties independent from our predictors, we will not see significant effects.

Model Validation Table 7 shows the results of multicollinearity analysis on the set of predictors. All VIF values are smaller than 2, which indicates the absence of problematic multicollinearity.

5.4 Results

Table 8 shows the performance of six resolvers on the complete GAP development set (overall and separately for Male and Female). It probably does not come as a surprise that performance increases over time; it is positive to note, though, that the Bias decreases correspondingly.

Table 9 shows the main results of our regression analysis on the subset of the GAP development set with a correct solution (cf. Section 5.1), organized by columns (systems). Each row provides a regression coefficient with its confidence level as well as the relative importance score for the predictor, using Dominance Analysis

		Lee et al. (2013)	Clark and Manning (2015)	Wiseman et al. (2016)	Lee et al. (2017)	Lee et al. (2018)	Joshi et al. (2020)
Gender	Coef	-0.473^{***}	-0.308^{**}	-0.314^{**}	-0.271^{**}	-0.215^{*}	-0.084
	DA	0.008	0.004	0.004	0.003	0.002	0.000
C_Freq	Coef	0.004	0.018^{***}	-0.004	-0.003	-0.001	0.001
	DA	0.000	0.005	0.000	0.000	0.000	0.000
I_Freq	Coef	-0.003	-0.003	-0.004	-0.006	-0.003	0.003
	DA	0.000	0.000	0.000	0.001	0.000	0.000
C_Diff	Coef	1.291^{**}	-1.617^{***}	$-0.933^{\cdot}$	-0.337	0.608	-0.065
	DA	0.006	0.002	0.001	0.001	0.001	0.000
I_Diff	Coef	-1.027^{*}	1.444^{***}	-0.086	$-0.740^{\cdot}$	-0.510	-0.053
	DA	0.003	0.002	0.001	0.004	0.001	0.000
C_Single	Coef	0.344^{**}	0.475^{***}	0.775^{***}	0.666^{***}	0.554^{***}	0.171
	DA	0.004	0.008	0.021	0.016	0.010	0.001
I_Single	Coef	-0.053	$-0.166^{\cdot}$	-0.268^{**}	-0.346^{***}	-0.360^{***}	0.036
	DA	0.001	0.001	0.003	0.006	0.006	0.000
C_Same	Coef	-0.603^{***}	-0.456^{***}	-0.561^{***}	-0.564^{***}	-0.336^{*}	-0.007
	DA	0.015	0.002	0.007	0.008	0.004	0.000
I_Same	Coef	0.086	0.366^{**}	0.120	0.318^{**}	0.317^{**}	0.028
	DA	0.000	0.002	0.000	0.003	0.003	0.000
Model Fit	R^2_{VZ}	0.05	0.04	0.05	0.05	0.03	0.01
	Acc	0.61 (0.58)	0.57 (0.55)	0.58 (0.53)	0.59 (0.53)	0.63 (0.63)	0.55 (0.55)

Table 9: Regression-based analysis of coreference resolution systems on GAP dataset. DA: Dominance Analysis, Freq: Frequency, C_: Correct, I_: Incorrect instances.

(DA). R^2_{VZ} indicates the goodness of fit values at the level of complete systems. (Note that these numbers, computed for logistic regression models, are not comparable to the numbers for linear regression models from Experiment 1.)

We also report accuracy values for the predictions of our logistic regression model, averaged over 10-fold cross-validation (*Acc*). Numbers in parentheses indicate the accuracy of corresponding majority baselines. The differences in baseline scores across systems are due to the fact that gold labels (i.e., the p(Correct) variable in the equation) are dependent on system predictions.

System level analysis We first discuss results at the system level. The last row of Table 9 (Model Fit) shows the overall model fit for all systems. The ability of our regression model to outperform majority baselines for the first four systems (Lee et al., 2013; Clark and Manning, 2015; Wiseman et al., 2016; Lee et al., 2017) shows that our analysis can predict mistakes made by these coreference resolvers by only considering a small set of discourse-related features plus Gender. In contrast, Lee et al. (2018) and Joshi et al. (2020) both show an R^2_{VZ} of almost zero, that is, the logistic regression models perform at the level of a majority class baseline – the remaining errors that they systems make are idiosyncratic rather than systematic. These findings tie in well with the overall system performance scores shown in Table 8.

It is striking that Joshi et al. (2020), the best model by a substantial margin, is also the one exhibiting the smallest bias. We see two possible explanations: (a), the model was trained on a large corpus from several domains with different discourse style, which may make it more robust to gender bias (Saunders and Byrne, 2020); (b) in contrast to the older studies, this model is based on contextualized embeddings, which also showed lower bias in Experiment 1. Without re-training the model, we cannot currently distinguish between these two explanations.

Predictor level analysis We now move on to investigate the contribution of each predictor to the systems' predictability. At this level, gender is a statistically significant predictor ($p < 0.05$) for all systems except Joshi et al. (2020). It has a negative sign throughout, indicating worse performance for female entities. This is again in line with the findings reported in Table 8. However, our approach reveals other important patterns which cannot be observed by using traditional analysis methods. First, Clark and Manning (2015) and Wiseman et al. (2016) have the same DA coefficient for gender variable but different R^2_{VZ} values. We interpret this to mean that the contribution of gender bias to the overall bias in these two systems is not the same, an observation that would not have been possible through traditional bias analysis methods (cf. Table 8).

Second, we see that the coefficient signs of the predictors C_Single and C_Same remain the same across systems: Systems perform better for instances where the correct antecedent is a single word, and it is not in the same sentence with the pronoun. Moreover, dominance analysis shows that these two predictors are among the main contributors to the biased predictions in four systems out of six, the two exceptions being Lee et al. (2013) and Joshi et al. (2020).

Third, the small but consistent positive relative importance values of the C_Diff and I_Diff predictors for half of the systems show that these variables help explain the systems' predictions. In contrast, the low relative importance values of the C_Frequency and I_Frequency predictors indicate that these variables do not affect coreference resolution much.

Interpretability These detailed findings indicate that, similar to emotion intensity prediction, the analysis of coreference resolvers can also benefit from not only the controlled bias variable but also from other properties of the input even in datasets which are designed carefully to isolate the effect of the target variable. As stated in Exp. 1, these analyses can also be used to extract interesting examples and subsets.

We illustrate this for the two attributes C_Same and I_Same, i.e., whether the correct and incorrect antecedent are in the same sentence or not. We split the GAP dataset into four reasonably-sized subsets based on the values of these attributes: the subset where both are in the same sentence (C_Same=1 and I_Same=1) includes ∼ 900 examples and the other three subsets include ∼ 300 examples. Table 10 shows the bias values (defined as above) for the three best performing systems. We observe that these systems vary widely regarding the subset where gender bias is most prominently visible varies across systems: Lee et al. (2017, 2018) both show the worst bias when the incorrect antecedent is not in the current sentence (I_Same=0), but differ in the effect of the position of the correct antecedent (C_Same). In contrast, Joshi et al. (2020) performs almost perfectly when I_Same=0, but struggles most the case when both correct and incorrect antecedent are in the current sentence. These variations in model performance across subsets raise questions about the representations of antecedents in the various models which go beyond the scope of this paper.

		I_Same=0	I_Same=1
Lee et al. (2017)	C_Same=0	**0.80**	1.10
	C_Same=1	0.90	0.90
		I_Same=0	I_Same=1
Lee et al. (2018)	C_Same=0	0.90	1.00
	C_Same=1	**0.86**	0.97
		I_Same=0	I_Same=1
Joshi et al. (2020)	C_Same=0	1.02	1.02
	C_Same=1	0.99	**0.94**

Table 10: Bias values for the three best performing systems, with data split into four groups according to C_Same and I_Same (worst bias marked in boldface).

6 Conclusion

In this article, we have argued that bias analysis, a task of major importance concerning the societal implications of NLP, can benefit from richer statistical methods to detect, quantify and attribute bias. We have proposed to follow other scientific fields in adopting regression analysis which (a) generalizes to multiple bias variables, (b) can quantify the contribution of confounder variables to the observed bias with measures of effect size, and (c) can be used to diagnose system behavior and extract informative datapoints.

Clearly, regression analysis is no panacea on its own: it presupposes a set of plausible covariates of bias, which can come from a wide variety of sources, including task-specific annotation, task-unspecific input representations, or model architecture (Hovy and Prabhumoye, 2021). Such covariates are typically known through domain expertise or uncovered by exploratory data analysis. Furthermore, the values of these bias variables must be available, or annotated, for all data points, which can represent a bottleneck. Thus, regression analysis complements, but does not replace, traditional methods of bias analysis.

We have demonstrated the usefulness of our approach by analyzing a range of model architectures on a regression task and a classification task, obtaining model-level results that are in line with the existing literature, e.g., BERT-based systems appear to exhibit comparatively little bias (Basta et al., 2019). In addition, adding predictor-level analysis offers a richer understanding of the importance of the bias variables and their interactions with other textual properties. Note that we only considered datasets specifically designed to exhibit the effects of a single bias variable. We believe that the benefits of our analysis framework would be even clearer on more naturalistic datasets where pairwise hypothesis tests become even more problematic (see, e.g., Gorrostieta et al. (2019)).

Another methodological debate that we hope to contribute to is what constitutes a 'substantial' bias? We have argued that effect sizes offer a statistically sound approach to measuring the amount of variation in the output that can be attributes to a set of input properties. Our study provides a starting point for the community to establish a magnitude for what it considers a 'substantial' bias, similar to the often-used thresholds for inter-annotator agreement (Cohen, 1968) or general effect sizes in psychology (Cohen, 1988).

Regarding future work, one avenue concerns richer regression models that analyze interactions among predictors. Such interactions, when properly motivated, can further improve our understanding of the performance data. In fact, our last example in Exp. 2 essentially demonstrates an interaction: the degree of gender bias in the conference resolvers is affected by an interaction between the position of the incorrect and the correct antecedents. Ideally, such observations might serve as motivation for assessing and potentially modifying model architectures or training regimens.

Another avenue of future research is widening our scope from the analysis of bias in NLP models (that is, "in vitro" bias according to our terminology in Section 1) to real life "in vivo" bias in academic communities. Recent studies have identified multiple such biases, e.g., gender bias in publications (Mohammad, 2020) and hiring (Eaton et al., 2020). We would hope that the application of robust regression analysis, a standard method in the social sciences, would help bolstering these studies and contribute towards redressing such social harms.

References

Achen, Christopher H. 1982. *Interpreting and using regression*, volume 29 of *Quantitative Applications in the Social Sciences.* Sage.

Aono, Masaki and Shinnosuke Himeno. 2018. KDE-AFFECT at SemEval-2018 Task 1: Estimation of affects in tweet by using convolutional neural network for n-gram. In *Proceedings of SemEval*, pages 156–161, New Orleans, LA.

Azen, Razia and Nicole Traxel. 2009. Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics*, 34:319 –347.

Baayen, Harald. 2008. *Analyzing Linguistic Data.* Cambridge University Press.

Barr, Dale J, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

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

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen. 2018. Parsimonious mixed models. ArXiv preprint, `http://arxiv.org/abs/1506.04967`.

Beinborn, Lisa, Torsten Zesch, and Iryna Gurevych. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, 2:517–530.

Bender, Emily M 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.

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

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of NeurIPS*, pages 4349–4357.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Kraemer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Science.

Budescu, David V. 1993. Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin*, 114(3):542.

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

Caucheteux, Charlotte and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134.

Clark, Kevin and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. 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)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.

Cohen, Jacob. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

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

Dayanik, Erenay and Sebastian Padó. 2020. Masking actor information leads to fairer political claims detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4385–4391, Online. Association for Computational Linguistics.

Dev, Sunipa, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7659–7666. AAAI Press.

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

Díaz, Mark, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Doll, Richard, Richard Peto, Jillian Boreham, and Isabelle Sutherland. 2004. Mortality in relation to smoking: 50 years' observations on male british doctors. *BMJ*, 328(7455):1519.

Dormann, Carsten, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, T. Diekötter, Jaime García Márquez, Bernd Gruber, Bruno Lafourcade, Pedro Leitão, Tamara Münkemüller, Colin Mcclean, Patrick Osborne, Björn Reineking, Boris Schröder, Andrew Skidmore, Damaris Zurell, and Sven Lautenbach. 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36:27–46.

Eaton, Asia A., Jessica F. Saunders, Ryan K. Jacobson, and Keon West. 2020. How gender and race stereotypes impact the advancement of scholars in stem: Professors'biased evaluations of physics and biology post-doctoral candidates. *Sex Roles*, 82(3):127–141.

Elazar, Yanai 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.

Feder, Amir, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Friedman, Batya and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347.

Gao, Yang, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. 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 1274–1290, Minneapolis, Minnesota. Association for Computational Linguistics.

Gardner, M. J. 1973. Using the environment to explain and predict mortality. *Journal of the Royal Statistical Society. Series A (General)*, 136(3):421–440.

Garg, Nikhil, 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.

Gerz, Daniela, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of EMNLP*, pages 316–327, Brussels, Belgium.

Gonen, Hila 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 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Gorrostieta, Cristina, Reza Lotfian, Kye Taylor, Richard Brutti, and John Kane. 2019. Gender de-biasing in speech emotion recognition. In *Proceedings of Interspeech*, pages 2823–2827.

Grömping, Ulrike. 2006. Relative importance for linear regression in R: the package `relaimpo`. *Journal of statistical software*, 17(1):1–27.

Hajat, Anjum, Charlene Hsia, and Marie S O'Neill. 2015. Socioeconomic disparities and air pollution exposure: a global review. *Current Environmental Health Reports*, 2(4):440–450.

Hovy, Dirk and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Imdad Ullah, Muhammad, Muhammad Aslam, Saima Altaf, and Munir Ahmed. 2019. Some new diagnostics of multicollinearity in linear regression model. *Sains Malaysiana*, 48(2):2051–2060.

Jaeger, T. Florian. 2008. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434–446. Special Issue: Emerging Data Analysis.

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

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Kaneko, Masahiro and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Kaneko, Masahiro and Danushka Bollegala. 2021a. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Kaneko, Masahiro and Danushka Bollegala. 2021b. Dictionary-based debiasing of pre-trained word embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 212–223, Online. Association for Computational Linguistics.

Kiritchenko, Svetlana and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of STARSEM*, pages 43–53, New Orleans, LA.

Kumar, Sachin, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. 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 4153–4163, Hong Kong, China. Association for Computational Linguistics.

Këpuska, V. and G. Bohouta. 2018. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 99–103.

Lauscher, Anne and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

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

Li, Shoushan, Rui Xia, Chengqing Zong, and Chu-Ren Huang. 2009. A framework of feature selection methods for text categorization. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 692–700, Suntec, Singapore. Association for Computational Linguistics.

Lindeman, Richard H., Peter F. Merenda, and Ruth Z. Gold. 1980. *Introduction to Bivariate and Multivariate Analysis*. Scott Foresman, Glenview, IL, USA.

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

McNamee, Roseanne. 2005. Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine*, 62(7):500–506.

Mehrabi, Ninareh, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2019. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM conference on Hypertext and Social Media*, pages 231–232.

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

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (Workshop Track)*.

Mohammad, Saif, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of SEMEVAL*, pages 1–17, New Orleans, LA.

Mohammad, Saif M. 2020. Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7860–7870, Online. Association for Computational Linguistics.

Mohammad, Saif M. and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the sixth joint conference on lexical and computational semantics (*Sem)*, Vancouver, Canada.

Padrão, Patrícia, Nuno Lunet, Ana Cristina Santos, and Henrique Barros. 2007. Smoking, alcohol, and dietary choices: evidence from the Portuguese national health survey. *BMC Public Health*, 7(1):1–9.

Papay, Sean, Roman Klinger, and Sebastian Padó. 2020. Dissecting span identification tasks with performance prediction. In *Proceedings of EMNLP*, page 4881–4895, Online.

Park, Ji Ho, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

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

Poliak, Adam, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Qian, Yusu, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

Roller, Stephen and Katrin Erk. 2016. PIC a different word: A simple model for lexical substitution in context. In *Proceedings of NAACL/HLT*, pages 1121–1126, San Diego, California.

Rubin, Donald B. 1973. Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

Rudinger, Rachel, 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*, pages 8–14, New Orleans, LA.

Salmerón, R., C. B. García, and J. García. 2018. Variance inflation factor and condition number in multiple linear regression. *Journal of Statistical Computation and Simulation*, 88(12):2365–2384.

Sap, Maarten, 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 Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Saunders, Danielle and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Schmid, Hans-Jörg. 2002. Do women and men really live in different cultures? Evidence from the BNC. In *Corpus linguistics by the Lune: a Festschrift for Geoffrey Leech*, pages 185–221. Peter Lang, Frankfurt.

Schuster, Mike and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Schwemmer, Carsten and Sebastian Jungkunz. 2019. Whose ideas are worth spreading? the representation of women and ethnic groups in ted talks. *Political Research Exchange*, 1(1):1–23.

Singh, Amit, Catherine Rose, Karthik Visweswariah, Vijil Chenthamarakshan, and Nandakishore Kambhatla. 2010. Prospect: A system for screening candidates for recruitment. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, page 659–668, New York, NY, USA. Association for Computing Machinery.

Smith, Thomas J. and C. M. Mckenna. 2013. A comparison of logistic regression Pseudo R^2 indices. *General Linear Model Journal*, 39(2):17–26.

Snijders, Tom and Roel Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd edition. Sage Publishers, London.

Stanovsky, Gabriel, 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.

Stuart, Elizabeth A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Sullivan, Gail M. and R. Feinn. 2012. Using effect size – or why the *p* value is not enough. *Journal of graduate medical education*, 4(3):279–82.

Sun, Dongming, Xiaolu Zhang, Kim-Kwang Raymond Choo, Liang Hu, and Feng Wang. 2021. Nlp-based digital forensic investigation platform for online communications. *Computers & Security*, 104:102210.

Sun, Tony, 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 Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Swinger, Nathaniel, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311.

Tjønneland, Anne, Morten Grønbæk, Connie Stripp, and Kim Overvad. 1999. Wine intake and diet in a random sample of 48763 Danish men and women. *The American journal of clinical nutrition*, 69(1):49–54.

Wang, Min and Xiaobing Zhou. 2018. Yuan at SemEval-2018 Task 1: Tweets emotion intensity prediction using ensemble recurrent neural network. In *Proceedings of SEMEVAL*, pages 205–209, New Orleans, LA.

Webster, Kellie, 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.

Wiseman, Sam, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, California. Association for Computational Linguistics.

Wu, Chuhan, Fangzhao Wu, Junxin Liu, Zhigang Yuan, Sixing Wu, and Yongfeng Huang. 2018. THU_NGN at SemEval-2018 Task 1: Fine-grained tweet sentiment intensity analysis with attention CNN-LSTM. In *Proceedings of SEMEVAL*, pages 186–192, New Orleans, Louisiana.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. `https://arxiv.org/abs/1609.08144`.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Zhao, Jieyu, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

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

Zhao, Jieyu, 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, Copenhagen, Denmark. Association for Computational Linguistics.

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

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015a. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Zhu, Zhiwei, Zhi Li, David Wylde, Michael Failor, and George Hrischenko. 2015b. Logistic regression for insured mortality experience studies. *North American Actuarial Journal*, 19(4):241–255.