

Multi-Dimensional Gender Bias Classification

Emily Dinan*, Angela Fan*†, Ledell Wu, Jason Weston, Douwe Kiela, Adina Williams

Facebook AI Research

†Laboratoire Lorrain d’Informatique et Applications (LORIA)

Abstract

Machine learning models are trained to find patterns in data. NLP models can inadvertently learn socially undesirable patterns when training on gender biased text. In this work, we propose a novel, general framework that decomposes gender bias in text along several pragmatic and semantic dimensions: bias from the gender of the person being spoken about, bias from the gender of the person being spoken to, and bias from the gender of the speaker. Using this fine-grained framework, we automatically annotate eight large scale datasets with gender information. In addition, we collect a new, crowdsourced evaluation benchmark. Distinguishing between gender bias along multiple dimensions enables us to train better and more fine-grained gender bias classifiers. We show our classifiers are valuable for a variety of applications, like controlling for gender bias in generative models, detecting gender bias in arbitrary text, and classifying text as offensive based on its genderedness.

1 Introduction

Language is a primary means by which people communicate, express their identities, and categorize themselves and others both explicitly and implicitly. Such social information is present in the words we write and, consequently, in the text we use to train our NLP models. In particular, models often can unwittingly learn negative associations about protected groups present in their training data and propagate them. In particular, NLP models often learn to replicate unwanted gender biases present in society (Bolukbasi et al., 2016; Hovy and Spruit, 2016; Caliskan et al., 2017; Rudinger et al., 2017; Garg et al., 2018; Gonen and Goldberg, 2019; Dinan et al., 2020). Since unwanted gender biases can affect downstream applications—sometimes even

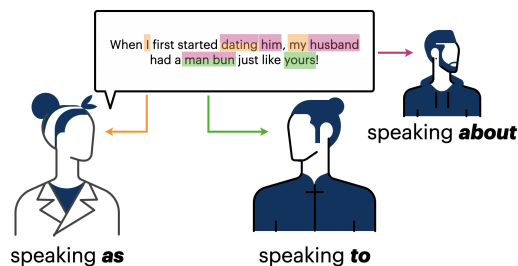


Figure 1: **Framework for Gender Bias in Dialogue.** We propose a framework separating gendered language based on who you are speaking ABOUT, speaking TO, and speaking AS.

leading to poor user experiences—understanding and mitigating gender bias is an important step towards making NLP tools and models safer and more equitable. In this paper, we provide a fine-grained framework for that purpose, analyze the presence of gender bias in models and data, and empower others by releasing tools that can be further applied to numerous text-based use-cases.

While many works have explored methods for removing gender bias from text (Emami et al., 2019; Hall Maudslay et al., 2019; Ravfogel et al., 2020), no extant work on classifying gender or removing gender bias has foregrounded the fact that we use language, at least in part, to collaboratively and socially construct our gender identities. We propose a pragmatic and semantic framework for measuring bias along three dimensions that is sensitive to conversational and performative aspects of gender, as illustrated in Figure 1. Recognizing these dimensions is important, because gender along each dimension can affect text differently, for example, by modifying word choice or imposing different preferences on sentence structure.

Decomposing gender into separate dimensions also allows for better identification of gender bias, which subsequently enables us to train a suite of classifiers for detecting different kinds of gender

*Joint first authors.

M	F	N
akin	feminist	optional
vain	lesbian	tropical
descriptive	uneven	volcanic
bench	transgender	glacial
sicilian	feminine	abundant

Table 1: **Bias in Wikipedia.** We compare the most over-represented adjectives in Wikipedia biographies of men and women to those in gender-neutral pages. We use a part-of-speech tagger (Honnibal and Montani, 2017), and computed $P(\text{word} \mid \text{gender})/P(\text{word})$ for words that appear more than 500 times.

bias in text. We train several classifiers on publicly available data that we annotate with gender information along our dimensions. We also collect a new crowdsourced dataset (MDGENDER) for better fine-grained evaluation of gender classifier performance. The classifiers we train have a wide variety of potential applications. We evaluate them on three: controlling the genderedness of generated text, detecting gender biased text, and examining the relationship between gender bias and offensive language. In addition, we expect these classifiers to be useful for future text applications such as detecting gender imbalance in newly created training corpora or model-generated text.

This paper makes four novel contributions: (i) we propose a multi-dimensional framework (ABOUT, AS, TO) for measuring and mitigating gender bias in language and NLP models, (ii) we introduce an evaluation dataset for performing gender identification that contains utterances re-written from the perspective of a specific gender along all three dimensions, (iii) we build a suite of classifiers capable of labeling gender in both a single and multitask set up, and finally (iv) we illustrate our classifiers’ utility for several downstream applications. All datasets, annotations, and classifiers will be released publicly to facilitate further research into the important problem of gender bias in text.

2 Related Work

Gender affects myriad aspects of NLP, including corpora, tasks, algorithms, and systems (Chang et al., 2019; Costa-jussà, 2019; Sun et al., 2019). For example, statistical gender biases are rampant in word embeddings (Jurgens et al., 2012; Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Zhao et al., 2018b; Basta et al., 2019; Chaloner and Maldonado, 2019; Du et al., 2019;

Ethayarajh et al., 2019; Gonen and Goldberg, 2019; Kaneko and Bollegala, 2019; Kurita et al., 2019; Zhao et al., 2019; Wang et al., 2020)—including multilingual ones (Escudé Font and Costa-jussà, 2019; Gonen et al., 2019; Zhou et al., 2019)—and affect a wide range of downstream tasks including coreference resolution (Zhao et al., 2018a; Cao and Daumé III, 2020; Emami et al., 2019), part-of-speech and dependency parsing (Garimella et al., 2019), language modeling (Qian et al., 2019; Nangia et al., 2020), appropriate turn-taking classification (Lepp, 2019), relation extraction (Gaut et al., 2020), identification of offensive content (Sharifirad and Matwin, 2019; Sharifirad et al., 2019), and machine translation (Stanovsky et al., 2019; Hovy et al., 2020).

For dialogue, gender biases in training corpora have been found to be amplified in machine learning models (Lee et al., 2019; Dinan et al., 2020; Liu et al., 2019). While many of the works cited above proposed methods of mitigating the unwanted effects of gender on text, Hall Maudslay et al. (2019), Liu et al. (2019), Zmigrod et al. (2019), and Dinan et al. (2020) in particular relied on counterfactual data to alter the training distribution to offset gender-based statistical imbalances (see §4.2 for more discussion of training set imbalances). Also relevant is Kang et al. (2019, PASTEL), which introduced a parallel style corpus and showed gains on style-transfer across binary genders.

Most relevant to this work, Sap et al. (2020) proposed a framework for modeling pragmatic aspects of many social biases in text. Our work and theirs focus on complementary aspects of a larger goal—namely, making NLP safe and inclusive for everyone—but the two approaches differ in several ways. We treat statistical gender bias in human or model generated text specifically, and in detail. Sap et al. (2020) proposed a different but compatible perspective, and aimed to situate gender bias within the broader landscape of negative stereotypes in social media text, an approach that can make parallels apparent across different kinds of harmful text. Moreover, they considered different pragmatic dimensions than we do: they targeted negatively stereotyped commonsense implications in arguably innocuous statements, whereas we investigate pragmatic dimensions that straightforwardly map to conversational roles (i.e., topics, addressees, and creators of text).

Finally, when investigating gender biases, one

cannot ignore the **intersectionality** of gender identities, i.e., when gender non-additively interacts with other identity characteristics. Negative gender stereotyping is known to be alternatively weakened or reinforced by the presence of social attributes like dialect (Tatman, 2017), class (Degaetano-Ortlieb, 2018) and race (Davis, 1981; Crenshaw, 1989). These differences have been found to affect gender classification in images (Buolamwini and Gebru, 2018), and also in sentences encoders (May et al., 2019). We acknowledge that these are crucial considerations, and intend to incorporate them in future work. For a thorough survey and a critical discussion of best practices for researching social “biases” in NLP, including and beyond gender, see Blodgett et al. (2020).

3 Dimensions of Gender Bias

Gender permeates language differently depending on the conversational role played by the people using that language (see Figure 1). We decompose gender bias along multiple dimensions: bias when speaking ABOUT someone, bias when speaking TO someone, and bias from speaking AS someone. This framework enables both finer-grained understanding of bias and better classification of gender’s effects on text from multiple domains.

Definition of Gender We annotate gender with four potential values: *masculine*, *feminine*, *neutral* and *unknown*. We take *neutral* to contain characters with either non-binary gender identity, or an identity which is unspecified for gender *by definition* (e.g. a talking tree). We include an *unknown* category for when the gender is genuinely not known.

Speaking About: Gender of the Topic. It’s well known that we change how we speak about others depending on who they are (Hymes, 1974; Rickford and McNair-Knox, 1994), and what their gender identity is (Lakoff, 1973; Eckert and McConnell-Ginet, 1992). For example, adjectives which describe women have been shown to differ from those used to describe men in numerous situations (Trix and Psenka, 2003; Gaucher et al., 2011; Moon, 2014; Hoyle et al., 2019), as do verbs that take nouns referring to men as opposed to women (Guerin, 1994; Hoyle et al., 2019).

Speaking To: Gender of the Addressee. People often adjust their speech based on who they are speaking with—their addressee(s)—to show solidarity with their audience or to express social dis-

tance (Wish et al., 1976; Bell, 1984; Hovy, 1987; Rickford and McNair-Knox, 1994; Bell and Johnson, 1997; Eckert and Rickford, 2001). We expect the addressee’s gender to affect, for example, how a man might communicate with another man about hair styles or beard hygiene. This exchange would probably differ if the man was communicating instead with a woman about the same topic.

Speaking As: Gender of the Speaker. People react to content differently depending on who created it. Like race, gender is often described as a “fundamental” category for self-identification and self-description (Banaji and Prentice, 1994, 315), with men, women, and non-binary people differing in how they actively create their own gender identities (West and Zimmerman, 1987). Who someone is *speaking as* strongly affects what they may say and how they say it, down to the level of their choices of adjectives and verbs in self-descriptions (Charyton and Snelbecker, 2007; Wetzal et al., 2012).

4 Creating Gender Classifiers

In an ideal world, we would expect little difference between texts describing men, women, and people with other gender identities. A machine learning model, then, would be unable to pick up on statistical differences in gendered language (i.e., statistical gender bias), because such differences would not exist. However, gender-based distributional differences *do* exist in current-day text (Table 1), and current-day gender bias classifiers *can* achieve much better than random performance (§5). Thus, we believe the aim of research like ours should be to work towards training the best and most sensitive gender classifier imaginable. If we had such an idealized classifier, it should eventually achieve random performance on future datasets, thereby signalling that we managed to create a dataset that is not gender biased. We take the classifiers we introduce here to be first steps towards this goal.

Previous work on gender bias classification has been predominantly single-task—often supervised on the task of analogy—and relied mainly on word lists, that are binarily gendered (Bolukbasi et al., 2016; Zhao et al., 2018b, 2019; Gonen and Goldberg, 2019)—sometimes also explicitly (Caliskan et al., 2017; Hoyle et al., 2019).¹ While wordlist-

¹We draw the distinction between **explicitly gendered** words, like *mother* and *father* which only refer to people of a specific gender, and **statistically gendered** words, like *doctor* and *nurse*, which can refer to people of any gender but

Dataset	M	F	N	U	Dim
<i>Training Data</i>					
Wikipedia	10M	1M	1M	-	ABOUT
Image Chat	39K	15K	154K	-	ABOUT
Funpedia	19K	3K	1K	-	ABOUT
Wizard	6K	1K	1K	-	ABOUT
Yelp	1M	1M	-	-	AS
ConvAI2	22K	22K	-	86K	AS
ConvAI2	22K	22K	-	86K	TO
OpenSub	149K	69K	-	131K	AS
OpenSub	95K	45K	-	209K	TO
LIGHT	13K	8K	-	83K	AS
LIGHT	13K	8K	-	83K	TO
<i>Evaluation Data</i>					
MDGENDER	384	401	-	-	ABOUT
MDGENDER	396	371	-	-	AS
MDGENDER	411	382	-	-	TO

Table 2: **Dataset Statistics.** Dataset size and dimension for the eight training datasets and MDGENDER.

based approaches provided a solid start, they ultimately prove insufficient. First, they conflate different conversational dimensions of gender bias, and are therefore unable to detect the subtle, but very well-described, pragmatic differences of interest here. Second, most existing gendered word lists are limited to explicitly binarily gendered words (e.g., *mom* vs. *dad*). Not only is binary gender completely inadequate for the task, but excluding statistically gendered words is problematic—because they are also strong anchors of gender stereotypes (Bolukbasi et al. 2016; Ethayarajh et al. 2019, i.a.).

Instead, we develop classifiers that decompose gender bias over sentences into semantic and/or pragmatic dimensions (*about/to/as*), including gender information that (i) falls outside the male-female binary, (ii) can be contextually determined, and (iii) is statistically as opposed to explicitly gendered. In the subsequent sections, we provide details regarding the annotation of data, and details for training these classifiers.

4.1 Data

In this section, we describe how we annotated our training data, including both the 8 existing datasets and our novel evaluation dataset, MDGENDER.

Annotation of Existing Datasets. We select a variety of existing datasets for training. Since one of our main contributions is a suite of open-source general-purpose gender bias classifiers, we selected datasets for training based on three criteria: inclu-

are statistically biased towards one gender.

sion of inferrable information about one or more of our dimensions, diversity in textual domain, and high quality, open-source data.

The datasets are: Wikipedia, Funpedia (a less formal version of Wikipedia) (Miller et al., 2017), Wizard of Wikipedia (knowledge-based conversation) (Dinan et al., 2019c), Yelp Reviews², ConvAI2 (chit-chat dialogue) (Dinan et al., 2019b), ImageChat (chit-chat dialogue about an image) (Shuster et al., 2018), OpenSubtitles (dialogue from movies) (Lison and Tiedemann, 2016), and LIGHT (chit-chat fantasy dialogue) (Urbanek et al., 2019). Table 2 presents dataset statistics.

Some of the datasets contain gender annotations provided by existing work. For example, classifiers trained for style transfer algorithms have previously annotated the gender of Yelp reviewers (Subramanian et al., 2018). In other datasets, we infer the gender labels. For example, in datasets where users are first assigned a *persona* to represent before chatting, often the gender of the persona is predetermined. In some cases gender annotations are not provided. In these cases, we sometimes impute the label if we are able to do so with high confidence. More details regarding how this was done can be found in §A.1.

Evaluation Dataset: MDGENDER. To make our classifiers reliable on all dimensions across multiple domains, we train on a variety of datasets. However, none of the existing data covers all three dimensions at the same time, and furthermore, many of the gender labels are noisy. To enable reliable evaluation, we collect a specialized corpus, MDGENDER, which acts as a gold-labeled dataset for the *masculine* and *feminine* classes.

First, we collect conversations between two speakers. Each speaker is provided with a persona description containing gender information, then tasked with adopting that persona and having a conversation. They are also provided with small sections of a biography from Wikipedia as the conversation topic. Using personas biographies to frame the conversation encourages crowdworkers to discuss ABOUT/TO/AS gender information.

To ensure there is ABOUT/TO/AS gender information contained in each utterance, we perform a second pass over the dataset. In this next phase, we ask a second set of annotators to rewrite each utterance to make it very clear that they are speaking ABOUT a man or a woman, speaking AS a man or a

²<https://yelp.com/dataset>

Model	About			To			As			All Avg.
	Avg.	M	F	Avg.	M	F	Avg.	M	F	
SingleTask ABOUT	70.43	63.54	77.31	44.44	36.25	52.62	67.75	69.19	66.31	60.87
SingleTask TO	50.12	99.74	0.5	49.39	95.38	3.4	50.41	100	0.81	49.97
SingleTask AS	46.97	51.3	42.4	57.27	67.15	47.38	78.21	70.71	85.71	60.82
MultiTask	62.59	64.32	60.85	78.25	73.24	83.25	72.15	66.67	77.63	67.13

Table 3: **Accuracy on the novel evaluation dataset MDGENDER** comparing single task classifiers to our multitask classifiers. We report accuracy on the *masculine* and the *feminine* classes, as well as the average of these two metrics. Finally, we report the average (of the M-F averages) across the three dimensions. MDGENDER was collected to enable evaluation on the *masculine* and *feminine* classes, for which much of the training data is noisy.

Model	Multitask Performance				
	M	F	N	Avg.	Dim.
Wikipedia	87.4	86.65	55.2	77.22	ABOUT
Image Chat	36.48	83.56	33.22	51.09	ABOUT
Funpedia	75.82	82.24	70.52	76.2	ABOUT
Wizard	64.51	83.33	81.82	76.55	ABOUT
Yelp	73.92	65.08	-	69.5	AS
ConvAI2	44	65.65	-	54.83	AS
ConvAI2	45.98	61.28	-	53.63	TO
OpenSubtitles	56.95	59.31	-	58.12	AS
OpenSubtitles	53.73	60.29	-	57.01	TO
LIGHT	51.57	65.72	-	58.65	AS
LIGHT	51.92	68.48	-	60.2	TO

Table 4: **Performance of the multitask model.** We evaluate the multitask model on the test sets from our training data. We report accuracy on each gold label—masculine, feminine, and neutral—and the average of the three. We do not report accuracy on imputed labels.

woman, and speaking TO a man or a woman. For example, given the utterance *Hey, how are you today? I just got off work*, a valid rewrite to make the utterance ABOUT a woman could be: *Hey, what’s up? I went for a coffee with my friend and her dog after work* as the *her* indicates a woman. Annotators are additionally asked to label how confident they are that someone else could predict the given gender label, allowing for flexibility between explicit genderedness (like the use of *he* or *she*) and statistical genderedness. An example instance of the task is shown in Table 10 and the interface is shown in §A.2 Figure 2. Additionally, we provide demographic information about the annotators for this task in §A.2.

4.2 Models

We outline how these classifiers are trained to predict gender bias along the three dimensions, provid-

ing details of the classifier architectures as well as how the data labels are used. In the single-task setting, we predict *masculine*, *feminine*, or *neutral* for each dimension – allowing the classifier to predict any of the three labels for the *unknown* category). To obtain a classifier capable of multitasking across the *about/to/as* dimensions, we train a model to score and rank a set of possible classes given textual input. For example, if given *Hey, John, I’m Jane!*, the model is trained to rank elements of both the sets {TO:masculine, TO:feminine, TO:neutral} and {AS:masculine, AS:feminine, AS:neutral} and produce appropriate labels TO:masculine and AS:feminine. Models are trained and evaluated on the annotated datasets as well as MDGENDER.

Model Architectures. For single task and multitask models, we use a pretrained Transformer-based (Vaswani et al., 2017) architecture. The model takes a text sequence (such as “John Doe was a man”) and a set of labels corresponding to the gender along a given dimension (such as {ABOUT:masculine, ABOUT:feminine, ABOUT:neutral}) as input; the model then ranks this set according to the textual input (as described), and outputs the top element from the set (such as ‘ABOUT:masculine’). More specifically, the Transformer provides representations for the textual input and set of classes. Classes are then scored (and ranked) by taking a dot product between the representations of the textual input and a given class, following the bi-encoder architecture (Humeau et al., 2019) trained with cross entropy. We use the same architecture and pre-training as in BERT-base (Devlin et al., 2019): specifically, the architecture is a 12 layer transformer encoder base with 12 attention heads and an embedding size of 768. We use ParLAI for model training (Miller et al., 2017). We will release all data and models.

Model	Performance			
	M	F	N	Avg.
MultiTask	87.4	86.65	55.2	77.22
Wikipedia Only	88.65	88.22	68.58	81.82
-gend words	86.94	74.62	74.33	78.63
-gend words and names	82.10	82.52	55.21	73.28

Table 5: **Ablation of gender classifiers on the Wikipedia test set.** We report the model accuracy on the masculine, feminine, and neutral classes, as well as the average accuracy across them. We train classifiers (1) on the entire text (2) after removing explicitly gendered words using a word list and (3) after removing gendered words and names. While removing gendered words and names makes classification more challenging, the model still obtains high accuracy.

5 Results

5.1 About/To/As Gender Classification

Quality of Classification Models. We compare models that classify along a single dimension compared to one that multitasks across all three, using MDGENDER to evaluate. We measure the percentage accuracy for masculine and feminine classes. (Recall, the MDGENDER does not contain *unknown* or *neutral* classes.) More details on this new dataset can be found in Section 4.1. Classifier results on MDGENDER are shown in Table 3.

We find that the multitask classifier has the best average performance across all dimensions, with a small hit to single-task performance in the ABOUT and AS dimensions. As expected, the single task models are unable to transfer to other dimensions: this clearly shows that gender information manifests differently along each dimension. Further, it demonstrates that solely using existing datasets is inadequate, as they do not contain labeled data along all three dimensions. Training for a single task allows models to specialize to detect and understand the nuances of text that indicate bias along one of the dimensions. However, in a multitask setting, models can learn to generalize to understand what language characterizes bias across multiple dimensions: we particularly see this manifest in the TO dimension.

Performance by Dataset. The gender classifiers along the TO, AS and ABOUT dimensions are trained on a variety of different existing datasets across multiple domains. We analyze which datasets are the most difficult to classify correctly

in Table 4. We find that ABOUT is the easiest dimension, particularly data from Wikipedia or based on Wikipedia, such as Funpedia and Wizard of Wikipedia, achieving almost 80% accuracy. This could be driven by the number of discussions about named entities, so classifying text ABOUT someone may be easier if a name is present.

The TO and AS directions are both more difficult, likely as they involve more context clues rather than relying on textual attributes and surface forms such as *she* and *he* to predict correctly. We find that generally the datasets have similar performance, except Yelp restaurant reviews, which has a higher accuracy (70%) on predicting AS.

Analysis of Classifier Performance. We break down choices made during training by comparing different models on Wikipedia (ABOUT dimension). We train with the variations of masking out gendered words and names. As gendered words and names are strongly correlated with gender, masking can force models into a more challenging but nuanced setting where they must learn to detect bias from the remaining text. We present the results in Table 5: masking out gendered words and names makes classification more challenging, but the model is still able to obtain high accuracy, indicating that gender bias is deeply embedded in the language used.

6 Applications

We demonstrate the broad utility of our multitask classifier by applying it to three different downstream applications. First, we show that we can use the classifier to control the genderedness of generated text. Next, we demonstrate its utility in biased text detection by applying it to Wikipedia to find the most gendered biographies. Finally, we evaluate our classifier on offensive text to explore the interplay between offensive text and gender.

6.1 Controllable Generation

By learning to associate control variables with textual properties, generative models can be controlled at inference time to adjust the generated text based on the desired properties of the user. This has been applied to a variety of different cases, including generating text of different lengths (Fan et al., 2018a), generating questions in chit-chat (See et al., 2019), and reducing bias (Dinan et al., 2020).

Previous work in gender bias used word lists to control bias, but found that word lists were lim-

Generation Statistics		
Control Token	# Gend. words	Pct. masc.
TO:feminine	246	48.0
AS:feminine	227	51.0
ABOUT:feminine	1151	19.72
Word list, feminine	1158	18.22
TO:masculine	372	75.0
AS:masculine	402	71.6
ABOUT:masculine	800	91.62
Word list, masculine	1459	94.8

Table 6: **Word statistics** measured on text generated from 1000 different seed utterances from ConvAI2 for each control token. We measure the number of gendered words (from a word list) that appear in the generated text, and the percentage of masculine-gendered words among all gendered words.

ited in coverage and applicability to a variety of domains (Dinan et al., 2020). However, by decomposing bias along the TO, AS, and ABOUT dimensions, fine-grained control models can be trained to control these different dimensions separately. This is important in various applications — for example, one may want to train a chatbot with a specific personality, leaving the AS dimension untouched, but want the bot to speak to and about everyone in a similar way. In this application, we train three different generative models, each of which controls generation for gender along one of the TO, AS, and ABOUT dimensions.

Methods We generate training data by taking the multitask classifier and using it to classify 250,000 textual utterances from Reddit, using a previously existing dataset extracted and obtained by a third party and made available on pushshift.io. This dataset was chosen as it is conversational in nature, but not one of the datasets that the classifier was trained on. We then use the labels from the classifier to prepend the utterances with tokens that indicate its gender label along the dimension. For example for the ABOUT dimension, we prepend utterances with tokens *ABOUT:<gender_label>*. At inference time, we choose control tokens to manipulate the text generated by the model.

We also compare to a baseline for which the control tokens are determined by a word list: if an utterance contains more masculine-gendered words than feminine-gendered words from the word list it is labeled as *masculine* (and vice versa for *feminine*); if it contains no gendered words or an equal number of masculine and feminine gendered words, it

is labeled as *neutral*. Following Dinan et al. (2020), we combine several existing word lists (Zhao et al., 2018b, 2019; Hoyle et al., 2019).

For training, we fine-tune a large, Transformer sequence-to-sequence model pretrained on a Reddit dump made freely available by pushshift.io. At inference time, we generate text via top- k sampling (Fan et al., 2018b), with $k = 10$ with a beam size of 10, and 3-gram blocking. We force the model to generate a minimum of 20 BPE tokens.

Qualitative Results. Example generations from various methods are shown in Appendix Table 11. In these examples we see that controlling for gender along different dimensions yields highly varied responses, even for identical input. This illustrates why limiting control to word lists is not enough to capture different aspects of gender. For example, adjusting AS to ‘feminine’ causes the model to write *No, I’ve been working. I don’t think I can make friendships online*, whereas setting ABOUT to ‘feminine’ for the same exact input causes the model to write *I think the problem is she’s a girl, so there’s not a lot of opportunity to make friends*.

We can also see from these examples how the genderedness of text differs along each axis when we switch between conditioning on *masculine* and *feminine*. For example, adjusting AS to ‘feminine’ causes the model to write *Awww, that sounds wonderful*, whereas setting AS to masculine generates *You can do it bro!*

Quantitative Results. Quantitatively, we evaluate by generating 1000 utterances seeded from ConvAI2 using both *masculine* and *feminine* control tokens and counting the number of gendered words from a gendered word list that also appear in the generated text. Results are shown in Table 6.

Utterances generated using ABOUT control tokens contain many gendered words. One might expect this, as when one speaks ABOUT another person, one may refer to them using gendered pronouns. We observe that for the control tokens *TO:feminine* and *AS:feminine*, the utterances contain a roughly equal number of masculine-gendered and feminine-gendered words. This is a reflection of the distribution in the training data: the ConvAI2 and Opensubtitles data show similar trends: on the ConvAI2 data, fewer than half of the gendered words in *SELF:feminine* utterances are feminine-gendered, and on the Opensubtitles data, the ratio

Masculine genderedness scores		
Biographies	Average	Median
All	0.74	0.98
Men	0.90	0.99
Women	0.042	0.00085

Table 7: **Masculine genderedness scores of Wikipedia bios.** We calculate a *masculine genderedness score* for a Wikipedia page by taking the median $p_x = P(x \in \text{ABOUT:masculine})$ among all paragraphs x in the page, where P is the probability distribution given by the classifier. We report the average and median scores for all biographies, as well as for biographies of men and women respectively.

Percentage of masculine-gendered text				
Dim	Safe	Offensive	<i>t</i> -statistic	<i>p</i> -value
ABOUT	81.03	70.66	5.49	5.19e-08
TO	44.68	60.15	-22.02	1.94e-46
AS	42.29	65.12	-14.56	1.05e-99

Table 8: **Genderedness of offensive content.** We measure the percentage of utterances in both the "safe" and "offensive" classes that are classified as *masculine-gendered*, among utterances that are classified as either *masculine-* or *feminine-gendered*. We test the hypothesis that safe and offensive classes distributions of *masculine-gendered* utterances differ using a *t*-test and report the *p*-value for each dimension.

drops to one-third.³

6.2 Bias Detection

Creating classifiers along different dimensions can be used to detect gender bias in any form of text, beyond dialogue itself. Such methods are useful in applications such as detecting, removing, and rewriting biased writing. We investigate using the trained classifiers to detect the most gendered biographies in Wikipedia.

Methods. We apply the multitask model to detect the most gendered masculine and feminine biographies in Wikipedia. We calculate the probability of being *masculine* in the ABOUT dimension for each paragraph among 65,000 biographies. We calculate a *masculine genderedness score* for the page by taking the median amongst all its paragraphs.

Quantitative Results. We report the average and median *masculine genderedness scores* for all bi-

ographies in the set of 65,000 that fit this criteria Table 7. We observe that while on average, the biographies skew largely toward *masculine* (the average score is 0.74), the classifier is more confident in the *femininity* of pages about women than it is in the *masculinity* of pages about men: the average *feminine genderedness score* for pages about women is $1 - 0.042 = 0.958$, while the average *masculine genderedness score* for pages about men is 0.90. This might suggest that biographies about women contain more gendered text on average.

Qualitative Results. We show the pages with the minimum score (most feminine-gendered biographies) and the maximum score (most masculine-gendered biographies) in Table 12 in the Appendix. The most masculine-gendered biographies are mostly composers and conductors, likely due to the historical imbalance in these occupations. Amongst the most feminine gendered biographies, there are many actresses from the mid-20th century. By examining the *most feminine* gendered paragraphs, anecdotally we find these are often describing the subject’s life after retirement. For example, Linda Darnell’s biography contains the line *Because of her then-husband, Philip Liebmann, Darnell put her career on a hiatus*, which clearly reflects negative societal stereotypes about the importance of women’s careers (Hiller and Philliber, 1982; Duxbury and Higgins, 1991; Pavalko and Elder Jr., 1993; Byrne and Barling, 2017; Reid, 2018).

6.3 Offensive Content

Finally, the interplay and correlation between gendered text and offensive text is an important area for study, as many examples of explicitly or contextually gendered text are disparaging or have negative connotations (e.g., “cat fight” and “doll”). Particularly for dialogue, neither form is desirable in the output of any chatbot system. There is a growing body of research on detecting offensive language in text. In particular, there has been recent work aimed at improving the detection of offensive language in the context of dialogue (Dinan et al., 2019a). We investigate this relationship by examining the distribution of labels output by our gender classifier on data that is labeled for offensiveness.

Methods. We use the *Standard* training and evaluation dataset from Dinan et al. (2019a). We examine the relationship between genderedness and

³The Opensubtitles data recalls the Bechdel test, which asks “whether a work [of fiction] features at least two women who talk to each other about something other than a man.” (Wikipedia contributors, 2020)

offensive utterances by labeling the gender of utterances (along the three dimensions) in the “safe” and “offensive” classes using our multitask classifier. We then measure the ratio of utterances labeled as *masculine-gendered* among utterances labeled as either *masculine-* or *feminine-gendered*.

Quantitative Results. Results are shown in Table 8. We observe that, on the ABOUT dimension, both the safe data and offensive data are more likely to be *masculine* than *feminine*; however, the offensive data is relatively less likely to be *masculine*. On the other hand, on the AS and TO dimensions, the safe data is more likely to be labeled as *feminine* and the offensive data is more likely to be labeled as *masculine*. We test the hypothesis that these distributions are unequal using a T-test, and find that these results are significant.

Qualitative Results. To explore how offensive content differs when it is ABOUT women and ABOUT men, we identified utterances for which the model had high confidence (probability > 0.70) that the utterance was *feminine* or *masculine* along the ABOUT dimension. After excluding stop words and words shorter than three characters, we hand-annotated the top 20 most frequent words as being *explicitly gendered*, a *swear word*, and/or bearing *sexual connotation*. For words classified as *masculine*, 25% of words fell into these categories, whereas for words classified as *feminine*, 75% of the words fell into these categories.

7 Conclusion

We propose a general framework for analyzing gender bias along three dimensions: (1) gender of the person being spoken about (ABOUT), (2) gender of the addressee (TO), and (3) gender of the speaker (AS). We annotate eight large existing datasets and create an evaluation dataset for the task of detect bias along each of these dimensions. We train classifiers (single and multitask) that demonstrate their broad utility by displaying strong performance in controlling bias in dialogue, detecting genderedness in text such as Wikipedia, and highlighting gender differences in offensive text.

Acknowledgments

Thanks to Isabelle Kloumann and our other colleagues from Responsible AI at Facebook, the audience of the NY NLP Meetup, Isaac Bleaman,

Sarah Phillips, and Zeerak Waseem Butt for discussions on this topic.

References

- David Bamman and Noah A. Smith. 2014. [Unsupervised discovery of biographical structure from text](#). *Transactions of the Association for Computational Linguistics*, 2:363–376.
- Mahzarin R. Banaji and Deborah A. Prentice. 1994. The self in social contexts. *Annual review of psychology*, 45(1):297–332.
- Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. [Evaluating the underlying gender bias in contextualized word embeddings](#). In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.
- Allan Bell. 1984. Language style as audience design. *Language in society*, 13(2):145–204.
- Allan Bell and Gary Johnson. 1997. Towards a sociolinguistics of style. *University of Pennsylvania Working Papers in Linguistics*, 4(1):2.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. [Language \(technology\) is power: A critical survey of “bias” in NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.
- Joy Buolamwini and Timnit Gebru. 2018. [Gender shades: Intersectional accuracy disparities in commercial gender classification](#). In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA. PMLR.
- Alyson Byrne and Julian Barling. 2017. When she brings home the job status: Wives’ job status, status leakage, and marital instability. *Organization Science*, 28(2):177–192.
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Yang Trista Cao and Hal Daumé III. 2020. [Toward gender-inclusive coreference resolution](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

- Kaytlin Chaloner and Alfredo Maldonado. 2019. [Measuring gender bias in word embeddings across domains and discovering new gender bias word categories](#). In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.
- Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. 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): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.
- Christine Charyton and Glenn E Snelbecker. 2007. Engineers’ and musicians’ choices of self-descriptive adjectives as potential indicators of creativity by gender and domain. *Psychology of Aesthetics, creativity, and the arts*, 1(2):91.
- Marta R Costa-jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, pages 1–2.
- Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139.
- Angela Y Davis. 1981. *Women, race, & class*. Vintage.
- Stefania Degaetano-Ortlieb. 2018. [Stylistic variation over 200 years of court proceedings according to gender and social class](#). In *Proceedings of the Second Workshop on Stylistic Variation*, pages 1–10, New Orleans. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019a. [Build it break it fix it for dialogue safety: Robustness from adversarial human attack](#). 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 4537–4546, Hong Kong, China. Association for Computational Linguistics.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019b. The second conversational intelligence challenge (ConvAI2). *arXiv preprint arXiv:1902.00098*.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019c. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Yupei Du, Yuanbin Wu, and Man Lan. 2019. [Exploring human gender stereotypes with word association test](#). 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 6133–6143, Hong Kong, China. Association for Computational Linguistics.
- Linda E Duxbury and Christopher A Higgins. 1991. Gender differences in work-family conflict. *Journal of applied psychology*, 76(1):60.
- Penelope Eckert and Sally McConnell-Ginet. 1992. Communities of practice: Where language, gender and power all live. In *Locating power: Proceedings of the second Berkeley women and language conference*, volume 1, pages 89–99. Berkeley, CA: Berkeley University.
- Penelope Eckert and John R Rickford. 2001. *Style and sociolinguistic variation*. Cambridge University Press.
- Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. [The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics.
- Joel Escudé Font and Marta R. Costa-jussà. 2019. [Equalizing gender bias in neural machine translation with word embeddings techniques](#). In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.
- Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. [Understanding undesirable word embedding associations](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.
- Angela Fan, David Grangier, and Michael Auli. 2018a. [Controllable abstractive summarization](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne,

- Australia. Association for Computational Linguistics.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018b. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.
- Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. [Women’s syntactic resilience and men’s grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.
- Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109.
- Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. [Towards understanding gender bias in relation extraction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online. Association for Computational Linguistics.
- Hila Gonen and Yoav Goldberg. 2019. [Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hila Gonen, Yova Kementchedjhiava, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 2019 Workshop on Widening NLP*, pages 64–67, Florence, Italy. Association for Computational Linguistics.
- Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 165–174.
- Bernard Guerin. 1994. Gender bias in the abstractness of verbs and adjectives. *The Journal of social psychology*, 134(4):421–428.
- Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. [It’s all in the name: Mitigating gender bias with name-based counterfactual data substitution](#). 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 5267–5275, Hong Kong, China. Association for Computational Linguistics.
- Dana V Hiller and William W Philliber. 1982. Predicting marital and career success among dual-worker couples. *Journal of Marriage and the Family*, pages 53–62.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. [“you sound just like your father” commercial machine translation systems include stylistic biases](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.
- Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.
- Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.
- Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. [Unsupervised discovery of gendered language through latent-variable modeling](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy. Association for Computational Linguistics.
- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. [Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring](#). *arXiv preprint arXiv:1905.01969*.
- Dell Hymes. 1974. Ways of speaking. In R. Bauman and J. Sherzer, editors, *Explorations in the ethnography of speaking*, volume 1, pages 433–451. Cambridge: Cambridge University Press.
- David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. [SemEval-2012 task 2: Measuring degrees of relational similarity](#). In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages

- 356–364, Montréal, Canada. Association for Computational Linguistics.
- Masahiro Kaneko 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.
- Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. [\(male, bachelor\) and \(female, Ph.D\) have different connotations: Parallely annotated stylistic language dataset with multiple personas](#). 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 1696–1706, Hong Kong, China. Association for Computational Linguistics.
- Maximilian Klein, Harsh Gupta, Vivek Rai, Piotr Konieczny, and Haiyi Zhu. 2016. Monitoring the gender gap with wikidata human gender indicators. In *Proceedings of the 12th International Symposium on Open Collaboration*, pages 1–9.
- Maximilian Klein and Piotr Konieczny. 2015. Wikipedia in the world of global gender inequality indices: What the biography gender gap is measuring. In *Proceedings of the 11th International Symposium on Open Collaboration*, pages 1–2.
- Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. [Measuring bias in contextualized word representations](#). In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.
- Robin Lakoff. 1973. Language and woman’s place. *Language in society*, 2(1):45–79.
- Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019. Exploring social bias in chatbots using stereotype knowledge. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 177–180, Florence, Italy. Association for Computational Linguistics.
- Haley Lepp. 2019. Pardon the interruption: Automatic analysis of gender and competitive turn-taking in united states supreme court hearings. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 143–145, Florence, Italy. Association for Computational Linguistics.
- Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles.
- Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? Towards fairness in dialogue systems. *CoRR*, abs/1910.10486.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. [On measuring social biases in sentence encoders](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. [ParlAI: A dialog research software platform](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.
- Rosamund Moon. 2014. From gorgeous to grumpy: adjectives, age and gender. *Gender & Language*, 8(1).
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. [Crows-pairs: A challenge dataset for measuring social biases in masked language models](#). *arXiv preprint arXiv:2010.00133*.
- Eliza K. Pavalko and Glen H. Elder Jr. 1993. Women behind the men: Variations in wives’ support of husbands’ careers. *Gender & Society*, 7(4):548–567.
- Yusu Qian, 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.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. [Null it out: Guarding protected attributes by iterative nullspace projection](#). *arXiv*.
- Joseph Reagle and Lauren Rhue. 2011. Gender bias in wikipedia and britannica. *International Journal of Communication*, 5:21.
- Erin M Reid. 2018. Straying from breadwinning: Status and money in men’s interpretations of their wives’ work arrangements. *Gender, Work & Organization*, 25(6):718–733.
- John R Rickford and Faye McNair-Knox. 1994. Addressee-and topic-influenced style shift: A quantitative sociolinguistic study. *Sociolinguistic perspectives on register*, pages 235–276.
- Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. [Social bias in elicited natural language inferences](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. [Social bias frames: Reasoning about social and power implications of language](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. [What makes a good conversation? how controllable attributes affect human judgments](#). 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 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.
- Sima Sharifirad, Alon Jacovi, Israel Bar Ilan Univesity, and Stan Matwin. 2019. Learning and understanding different categories of sexism using convolutional neural network’s filters. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 21–23.
- Sima Sharifirad and Stan Matwin. 2019. Using attention-based bidirectional lstm to identify different categories of offensive language directed toward female celebrities. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 46–48.
- Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Engaging image chat: Modeling personality in grounded dialogue. *arXiv preprint arXiv:1811.00945*.
- Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. [Evaluating gender bias in machine translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.
- Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc’Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *arXiv preprint arXiv:1811.00552*.
- Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. [Mitigating gender bias in natural language processing: Literature review](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.
- Rachael Tatman. 2017. [Gender and dialect bias in YouTube’s automatic captions](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.
- Frances Trix and Carolyn Psenka. 2003. Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society*, 14(2):191–220.
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. [Learning to speak and act in a fantasy text adventure game](#). 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 673–683, Hong Kong, China. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It’s a man’s wikipedia? assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*.
- Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5(1):5.
- Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. [Double-hard debias: Tailoring word embeddings for gender bias mitigation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5443–5453, Online. Association for Computational Linguistics.
- Candace West and Don H Zimmerman. 1987. Doing gender. *Gender & society*, 1(2):125–151.
- Eunike Wetzel, Benedikt Hell, and Katja Pässler. 2012. Comparison of different test construction strategies in the development of a gender fair interest inventory using verbs. *Journal of Career Assessment*, 20(1):88–104.
- Wikipedia contributors. 2020. [Bechdel test — Wikipedia, the free encyclopedia](#). [Online; accessed 3-April-2020].
- Myron Wish, Morton Deutsch, and Susan J Kaplan. 1976. Perceived dimensions of interpersonal relations. *Journal of Personality and social Psychology*, 33(4):409.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. [Gender bias in contextualized word embeddings](#). In

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. [Learning gender-neutral word embeddings](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. [Examining gender bias in languages with grammatical gender](#). In *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 5275–5283, Hong Kong, China. Association for Computational Linguistics.

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. [Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Appendices

A.1 Existing Data Annotation

Many of our annotated datasets contain cases where the ABOUT, AS, TO labels are not provided (i.e. *unknown*). For example, often we do not know the gender of the content creator for Wikipedia (i.e., the AS dimension is *unknown*). To retain such examples for training, we either impute the gender label or provide a label at random. We apply the imputation strategy for data for which the ABOUT label is *unknown* using a classifier trained only on other Wikipedia data for which this label *is* provided. Data without a TO or AS label was assigned one at random, choosing between *masculine* and *feminine* with equal probability. From epoch to epoch, we switch these arbitrarily assigned labels so that the model learns to label *unknown* examples as *masculine* or *feminine* with roughly equal probability. This label flipping allows us to retain greater quantities of data by preserving unknown samples. During training, we balance the data across the *masculine*, *feminine*, and *neutral* classes by up-sampling classes with fewer examples. We describe in more detail how each of the eight training datasets is annotated:

1. **Wikipedia** - to annotate ABOUT, we use a Wikipedia dump and extract biography pages. We identify biographies using named entity recognition applied to the title of the page (Honnibal and Montani, 2017). We label pages with a gender based on the number of gendered pronouns (*he* vs. *she* vs. *they*) and label each paragraph in the page with this label for the ABOUT dimension.⁴ Wikipedia is well known to have gender bias in equity of biographical coverage and lexical bias in noun references to women (Reagle and Rhue, 2011; Graells-Garrido et al., 2015; Wagner et al., 2015; Klein and Konieczny, 2015; Klein et al., 2016; Wagner et al., 2016), making it an interesting test bed for our investigation.
2. **Funpedia** - Funpedia (Miller et al., 2017) contains rephrased Wikipedia sentences in a more conversational way. We retain only biography related sentences and annotate similar to Wikipedia, to give ABOUT labels.

⁴This method of imputing gender is similar to the one used in Reagle and Rhue (2011, 1142) and Bamman and Smith (2014), except we also incorporate non-oppositional gender categories, and rely on basic counts without scaling.

3. **Wizard of Wikipedia** - Wizard of Wikipedia (Dinan et al., 2019c) contains two people discussing a topic in Wikipedia. We retain only the conversations on Wikipedia biographies and annotate to create ABOUT labels.
4. **ImageChat** - ImageChat (Shuster et al., 2018) contains conversations discussing the contents of an image. We use the Xu et al. image captioning system⁵ to identify the contents of an image and select gendered examples.
5. **Yelp** - we use the Yelp reviewer gender predictor developed by (Subramanian et al., 2018) and retain reviews for which the classifier is very confident – this creates labels for the content creator of the review (AS). We impute ABOUT labels on this dataset using a classifier trained on the datasets 1-4.
6. **ConvAI2** - ConvAI2 (Dinan et al., 2019b) contains persona-based conversations. Many personas contain sentences such as *I am a old woman* or *My name is Bob* which allows annotators to annotate the gender of the speaker (AS) and addressee (TO) with some confidence. Many of the personas have unknown gender. We impute ABOUT labels on this dataset using a classifier trained on the datasets 1-4.
7. **OpenSubtitles** - OpenSubtitles⁶ (Lison and Tiedemann, 2016) contains subtitles for movies in different languages. We retain English subtitles that contain a character name or identity. We annotate the character’s gender using gender kinship terms such as *daughter* and gender probability distribution calculated by counting the masculine and feminine names of baby names in the United States⁷. Using the character’s gender, we get labels for the AS dimension. We get labels for the TO dimension by taking the gender of the next character to speak if there is another utterance in the conversation; otherwise, we take the gender of the *last* character to speak. We impute ABOUT labels on this dataset using a classifier trained on the datasets 1-4.
8. **LIGHT** - LIGHT contains persona-based conversation. Similarly to ConvAI2, annotators

⁵<https://github.com/AaronCCWong/Show-Attend-and-Tell>

⁶<http://www.opensubtitles.org/>

⁷<https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data>

	Percent of total
Man	67.38
Woman	18.34
Non-binary	0.21
Prefer not to say	14.07

Table 9: Self-reported gender identities of annotators used to collect the new evaluation dataset MDGENDER. Annotators were given the option to not answer this question or to select “prefer not to say.”

labeled the gender of each persona (Dinan et al., 2020), giving us labels for the speaker (AS) and speaking partner (TO). We impute ABOUT labels on this dataset using a classifier trained on the datasets 1-4.

A.2 New Evaluation Dataset

The interface for our new evaluation dataset MDGENDER can be seen in Figure 2. Examples from the new dataset can be found in Table 10.

This dataset was collected using crowdworkers from Amazon’s Mechanical Turk. All workers are English-speaking and located in the United States. During the “re-write phase” (described in §4.1) crowdworkers were asked to provide their own gender identity if they were willing. Workers were given the option to not answer this question or to select “prefer not to say.” Results from this survey are shown in Table 9. For privacy reasons we do not associate the self-reported gender of the annotator with the labeled examples in the dataset and only report these statistics in aggregate. Over two thirds of annotators identified as men, which may introduce its own biases into the dataset.

A.3 Applications

Example generations for various control tokens, as well as for our word list baseline, are shown in Table 11. See §6.1 on Controllable Generation in the main paper for more details.

The top 10 most gendered Wikipedia biographies are shown in Table 12. See §6.2 on Detecting Bias in the main paper for more details.

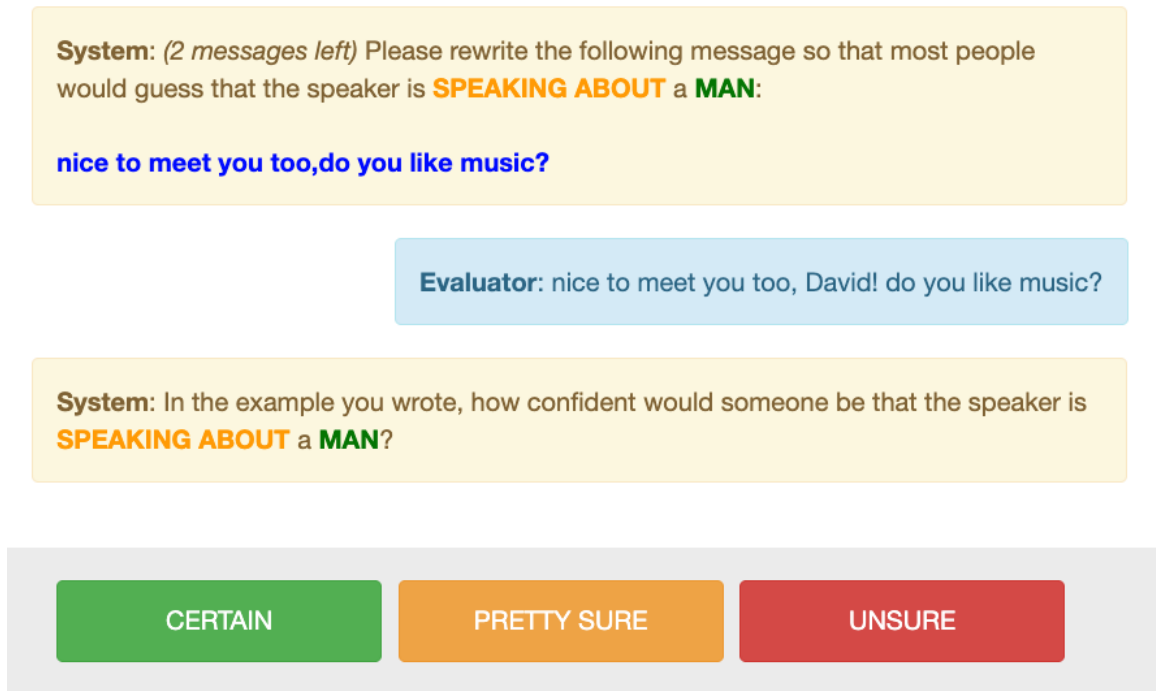


Figure 2: **Annotation interface.** Annotation interface for collecting MDGENDER. Annotators were shown an utterance from a conversation, and asked to re-write it such that it is clear they would be speaker about/to/as a man or a woman. They were then asked for their confidence level.

Utterance	Dim.	Gender	Confidence
<i>Original:</i> That's interesting. I am a Chef. What are your hobbies			
<i>Rewrite:</i> that's interesting. i am a chef and nfl player what are your hobbies	AS	masc.	certain
<i>Original:</i> My name is Zachery but most call me Z. What's your name miss?			
<i>Rewrite:</i> My name is Tina but most call me T. What's your name miss?	AS	fem.	pretty sure
<i>Original:</i> I said I like to wrestle with my kids for fun. What do you do for work?			
<i>Rewrite:</i> i said i like to wrestle with my son for fun. what do you do for work?	ABOUT	masc.	certain
<i>Original:</i> just checking up on you by saying hi			
<i>Rewrite:</i> Hanna said she was just checking up on you by saying hi	ABOUT	fem.	pretty sure
<i>Original:</i> How are you doing today Jahsim?			
<i>Rewrite:</i> How are you doing today Mr. Jahsim?	TO	masc.	certain
<i>Original:</i> My name is Aleksandro and I am a bee farmer!			
<i>Rewrite:</i> My name is alexsandro and I am a bee farmer! How about you ma'am?	AS	fem.	certain

Table 10: **Examples from the MDGENDER.** Crowdworkers were asked to re-write dialogue utterances such that most people would guess that the utterance was either said to, said by, or about a man or a woman. Afterwards, they were asked to give a confidence level in their re-write, meant to capture the differences between statistical biases (more men play football than women) and fact (you do not have to be a man to play football).

<p>Input: Awww! Cute! Our are a handful, but they are my babies! 4 wiener dogs. TO:feminine I can't wait for your children to get married.. I want one so badly! :) AS:feminine Awww, that sounds wonderful. Can't wait to see them in person! :) ABOUT:feminine You're my hero. That's a cute dog. How long has she been with you? Word list, feminine That's pretty awesome! She was a little bit overweight but she's a sweet girl now</p>
<p>Input: Does vegan mean you do not want to be a super hero? TO:feminine I'd love to be vegan, but it just isn't something I have any interest in doing. AS:feminine I don't want to become super heroes! I just think it would be cool to be vegan. ABOUT:feminine She's a girl. She can't be a superhero!!! 11! 1! 1 1! Word list, feminine It does. She is a vegan and I am a super - hero. I am not vegan, and I don't want to become any sort of superhero.</p>
<p>Input: That is an good. Maybe you can make friends in a club or on a job? TO:feminine I was going to do that, but then I realized I didn't want to be alone. AS:feminine No, I've been working. I don't think I can make friendships online. ABOUT:feminine I think the problem is she's a girl, so there's not a lot of opportunity to make friends. Word list, feminine I have a girlfriend and I live in a small town. I don't know what clubs she can make any friends in.</p>
<p>Input: Its been 2 years I am a grease monkey and been trying to stay busy TO:masculine I'm not sure this is true. I've seen a lot of grease monkeys on reddit. I don't think there's any evidence it's true. AS:masculine You can do it bro! Just keep at it! You got this!!! : d) ABOUT:masculine I'd love to be a grease guy but I just can't make it to the gym. : (Word list, masculine You're not alone bro. I'm trying to be that guy, but it's a little hard sometimes.</p>
<p>Input: Have you ever almost died? I nearly did from urinary cancer when I was 18. TO:masculine No one in my family has died. I have only seen a handful of people die. I am sure you will be fine. AS:masculine No I haven't. But it's a lot better now, thanks to all the help I've received. ABOUT:masculine He said * nearly *. It's more likely to happen when you're older than the average person. Word list, masculine I've only been alive for 3 days. I have no idea what I'm going to do next, but I'll be okay.</p>
<p>Input: I totally believe you. I love working out. TO:masculine I'm going to start doing this, thanks! If it works, I'll post it on r / fitness! AS:masculine Thanks! I've always been really into exercise, too, and it's definitely helped me. ABOUT:masculine Thank you! I'm really glad he does :). I've been working out since october. Word list, masculine Me too! I love being able to go to the gym without feeling like I'm just a little kid. It's so rewarding when you get back in the swing of things.</p>

Table 11: **Example generations** from a generative model trained using controllable generation, with control tokens determined by the classifier. Sequences are generated with top- k sampling, $k = 10$, with a beam size of 10 and 3-gram blocking. Input is randomly sampled from the ConvAI2 dataset.

<i>Most Feminine</i>	<i>Most Masculine</i>
1. Edie Sedgwick: was an American actress and fashion model...	1. Derek Jacobi: is an English actor and stage director...
2. Linda Darnell: was an American film actress...	2. Bohuslav Martinů: was a Czech composer of modern classical music...
3. Maureen O'Hara: was an Irish actress and singer...	3. Carlo Maria Giulini: was an Italian conductor...
4. Jessica Savitch: was an American television news presenter and correspondent,...	4. Zubin Mehta: is an Indian conductor of Western classical music...
5. Patsy Mink: Mink served in the U.S. House of Representatives...	5. John Barbirolli: was a British conductor and cellist ...
6. Shirley Chisholm: was an American politician, educator, and author...	6. Claudio Abbado: was an Italian conductor...
7. Mamie Van Doren: is an American actress, model, singer, and sex symbol who is...	7. Ed Harris: is an American actor, producer, director, and screenwriter...
8. Jacqueline Cochran: was a pioneer in the field of American aviation and one of t...	8. Richard Briers: was an English actor...
9. Chloë Sevigny: is an American actress, fashion designer, director, and form...	9. Artur Schnabel: was an Austrian classical pianist, who also composed and tau...
10. Hilda Solis: is an American politician and a member of the Los Angeles Co...	10. Charles Mackerras: was an Australian conductor...

Table 12: **Most gendered Wikipedia biographies** We ran our multitask classifier over 68 thousand biographies of Wikipedia. After selecting for biographies with a minimum number of paragraphs (resulting in 15.5 thousand biographies) we scored them to determine the most *masculine* and *feminine* gendered.