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 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 novel, crowdsourced evaluation benchmark of utterance-level gender rewrites. Distinguishing between gender bias along multiple dimensions is important, as it enables us to train finer-grained gender bias classifiers. We show our classifiers prove valuable for a variety of important applications, such as controlling for gender bias in generative models, detecting gender bias in arbitrary text, and shed light on offensive language in terms of genderedness.

1 Introduction

Language is a social behavior, and as such, it is a primary means by which people communicate, express their identities, and socially categorize themselves and others. 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 biases against others based on their gender (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., 2019a). Since unwanted gender biases can affect downstream applications—sometimes even leading to poor user experiences—understanding and mitigating gender bias is an important step towards making NLP tools and models safer, more equitable, and more fair. We provide a finer-grained framework for this purpose, analyze the presence of gender bias in models and data, and empower others by releasing tools that can be employed to address these issues for numerous text-based use-cases.

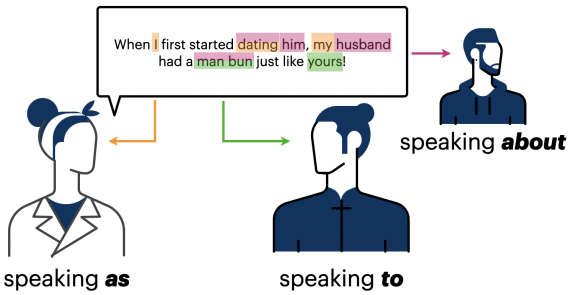

Figure 1: **Framework for Gender Bias in Dialogue**. We propose a framework separating gendered language based on who you are speaking ABOUT, speaking TO, and speaking AS.

While many works have explored methods for removing gender bias from text (Bolukbasi et al., 2016; Emami et al., 2019; Maudslay et al., 2019; Dinan et al., 2019a; Kaneko and Bollegala, 2019; Zmigrod et al., 2019; Ravfogel et al., 2020), no extant work on classifying gender or removing gender bias has incorporated facts about how humans collaboratively and socially construct our language and identities. We propose a pragmatic and semantic framework for measuring bias along three dimensions that builds on knowledge of the 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,

*Joint first authors.

by modifying word choice or imposing different preferences in how we construct sentences.

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 bias in text. We train several classifiers on freely available data that we annotate with gender information along our dimensions. We also collect a new crowdsourced dataset (MDGENDER) for better 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 gendered text, and examining the relationship between gender bias and offensive language. In addition, we expect them to be useful in future for many text applications such as detecting gender imbalance in newly created training corpora or model-generated text.

In this work, we make four main contributions: we propose a multi-dimensional framework (ABOUT, AS, TO) for measuring and mitigating gender bias in language and NLP models, 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, we train a suite of classifiers capable of labeling gender in both a single and multitask set up, and finally 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 language.

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; Gonen and Goldberg, 2019; Kaneko and Bollegala, 2019; Zhao et al., 2019)—even multilingual ones (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, 2019; Emami et al., 2019), part-of-speech and dependency parsing (Garimella et al., 2019), unigram language modeling (Qian et al., 2019), appropriate turn-taking classification (Lepp, 2019), relation extraction (Gaut et al., 2019), identification of offensive content (Sharifirad and Matwin, 2019; Sharifirad et al., 2019), and machine translation (Stanovsky et al., 2019). Furthermore, translations are judged as having been produced by older and more male speakers than the original was (Hovy et al., 2020).

For dialogue text particularly, gender biases in training corpora have been found to be amplified in machine learning models (Lee et al., 2019; Dinan et al., 2019a; Liu et al., 2019). While many of the works cited above propose methods of mitigating the unwanted effects of gender on text, Maudslay et al. (2019); Zmigrod et al. (2019); Dinan et al. (2019a) in particular rely on counterfactual data to alter the training distribution to offset gender-based statistical imbalances (see §4.1 for more discussion of training set imbalances). Also relevant is Kang et al. (2019, PASTEL), which introduces a parallel style corpus and shows gains on style-transfer across binary genders. In this work, we provide a clean new way to understand gender bias that extends to the dialogue use-case by independently investigating the contribution of author gender to data created by humans.

Most relevant to this work, Sap et al. (2019b) proposes a framework for modeling pragmatic aspects of many social biases in text, such as intent to offend, for guiding discovery of new instances of social bias. These works focus on complementary aspects of a larger goal—namely, making NLP safe and inclusive for everyone—but they differ in several ways. Here, we treat statistical gender bias in human or model generated text specifically, allotting it the focused and nuanced attention that such a complicated phenomenon deserves. Sap et al. (2019b) takes a different perspective, and aims to characterize the broader landscape of negative stereotypes in social media text, an approach which can make parallels apparent across different types of socially harmful content. Moreover, they consider different pragmatic dimensions than we do: they target negatively stereotyped commonsense implications in arguably innocuous statements, whereas we investigate pragmatic dimensions that straightforwardly map to conversational roles (i.e., topics, addressees, and authors of content). As such, we believe the two frameworks to be fully compatible.

Also relevant is the **intersectionality** of gender identity, i.e., when gender non-additively interacts with other identity characteristics. Negative gender stereotyping is known to be weakened or reinforced by the presence of other social factors, such as dialect (Tatman, 2017), class (Degaetano-Ortlieb, 2018) and race (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, but set them aside for follow-up work.

3 Dimensions of Gender Bias

Gender infiltrates language differently depending on the conversational role played by the people using that language (see Figure 1). We propose a framework for decomposing gender bias into three separate dimensions: bias when speaking ABOUT someone, bias when speaking TO someone, and bias from speaking AS someone. In this section, we first define *bias* and *gender*, and then motivate and describe our three dimensions.

3.1 Definitions of Bias and Gender

Bias. In an ideal world, we would expect little difference between texts describing men, women, and people with other gender identities, aside from the use of explicitly gendered words, like pronouns or names. A machine learning model, then, would be unable to pick up on statistical differences among gender labels (i.e., gender **bias**), because such differences would not exist. Unfortunately, we know this is not the case. For example, Table 1 provides examples of adjectives, adverbs, and verbs that are more common in Wikipedia biographies of people of certain genders. This list was generated by counting all verbs, adjectives, and adverbs (using a part-of-speech tagger from Honnibal and Montani (2017)) that appear in a large section of biographies of Wikipedia. We then computed $P(\text{word} \mid \text{gender})/P(\text{word})$ for words that appear more than 500 times. The top over-represented verbs, adjectives, and adverbs using this calculated metric are displayed for each gender.

In an imagined future, a classifier trained to identify gendered text would have (close to) random performance on non-gender-biased future data, because the future would be free of the statistical biases plaguing current-day data. These statistical biases are what make it possible for current-day classifiers to perform better than random chance. We know that current-day classifiers are gender biased, because they achieve much better than random performance by learning distributional differences in how current-day texts use gender; we show this in §5. These classifiers learn to pick up on these statistical biases in text *in addition to* explicit gender markers (like *she*).[1]

Gender. Gender manifests itself in language in numerous ways. In this work, we are interested in gender as it is used in English when referring to people and other sentient agents, or when discussing their identities, actions, or behaviors. We annotate gender with four potential values: *masculine, feminine, neutral* and *unknown* — which allows us to go beyond the oppositional male-female gender binary. We take the *neutral* category to contain characters with either non-binary gender identity, or an identity which is unspecified for gender by definition (say, for a magic tree).[2] We also include an *unknown* category for when there might be a gender identity at play, but the gender is not known or readily inferrable by crowdworkers from the text (e.g., in English, one would not be able to infer gender from just the short text "Hello!").

3.2 Gender in Multiple Dimensions

Exploring gender's influence on language has been a fruitful and active area of research in many disciplines, each of which brings its own unique perspectives to the topic (Lakoff, 1973; Butler, 1990; Cameron, 1990; Lakoff, 1990; Swann, 1992; Crawford, 1995; Weatherall, 2002; Sunderland, 2006; Eckert and McConnell-Ginet, 2013; Mills, 2014; Coates, 2015; Talbot, 2019). In this section, we propose a framework that decomposes gender's contribution along three conversational dimensions to enable finer-grained classification of gender's effects on text from multiple domains.

***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, in particular,

[1] We caution the reader that "the term bias is often used to refer to demographic disparities in algorithmic systems that are objectionable for societal reasons" (Barocas et al., 2020, 14); we restrict our use of **bias** to its traditional definition here.

[2] We fully acknowledge the existence and importance of all chosen gender identities—including, but not limited to non-binary, gender fluid, poly-gender, pan-gender, alia-gender, agender—for the end goal of achieving accessible, inclusive, and fair NLP. However, these topics require a more nuanced investigation than is feasible using naïve crowdworkers.

VERBS			ADJECTIVES			ADVERBS		
M	F	N	M	F	N	M	F	N
finance	steamed	increases	akin	feminist	optional	ethnically	romantically	westward
presiding	actor	range	vain	lesbian	tropical	intimately	aground	inland
oversee	kisses	dissipated	descriptive	uneven	volcanic	soundly	emotionally	low
survives	towed	vary	bench	transgender	glacial	upstairs	sexually	automatically
disagreed	guest	engined	sicilian	feminine	abundant	alongside	happily	typically
obliged	modelling	tailed	24-hour	female	variable	artistically	socially	faster
filling	cooking	excavated	optimistic	reproductive	malay	randomly	anymore	normally
reassigned	kissing	forested	weird	sexy	overhead	hotly	really	round
pledged	danced	upgraded	ordained	blonde	variant	lesser	positively	usually
agreeing	studies	electrified	factual	pregnant	sandy	convincingly	incredibly	slightly

Table 1: **Bias in Wikipedia**. We look at the most over-represented words in biographies of men and women, respectively, in Wikipedia. We also compare to a set of over-represented words in gender-neutral pages. We use a part-of-speech tagger (Honnibal and Montani, 2017) and limit our analysis to words that appear at least 500 times.

based on their gender (Lakoff, 1973). People often change how they refer to others depending on the gender identity of the individual being spoken about (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). Furthermore, metaphorical extensions—which can shed light on how we construct conceptual categories (Lakoff and Johnson, 1980)—to men and women starkly differ (Fontecha and Catalan 2003; Holmes 2013, 325; Amery et al. 2015).

***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 express social distance (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, the way a man might communicate with another man about styling their hair would differ from how he might communicate with a woman about the same topic.

***Speaking As*: Gender of the Speaker.** People react to content differently depending on who created it.[3] For example, Sap et al. (2019a) find that naïve annotators are much less likely to flag as offensive certain content referring to race, if they have been told the author of that content speaks a dialect that signals in-group membership (i.e., is less likely to be intended to offend). 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 and perceive of their own gender identities (West and Zimmerman, 1987). Who someone is *speaking as* strongly affect 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; Wetzel et al., 2012). Even children as young as two dislike when adults misattribute a gender to them (Money and Ehrhardt, 1972; Bussey, 1986), suggesting that gender is indeed an important component of identity.

Our *Speaking As* dimension builds on prior work on author attribution, a concept purported to hail from English logician Augustus de Morgan (Mendenhall, 1887), who suggested that authors could be distinguished based on the average word length of their texts. Since then, sample statistics and NLP tools have been used for applications such as settling authorship disputes (Mosteller and Wallace, 1984), forensic investigations (Frantzeskou et al., 2006; Rocha et al., 2016; Peng et al., 2016), or extracting a stylistic fingerprint from text that enables the author to be identified (Stamatatos et al., 1999; Luyckx and Daelemans, 2008; Argamon et al., 2009; Stamatatos, 2009; Raghavan et al., 2010; Cheng et al., 2011; Stamatatos, 2017). More specifically, automatic gender attribution has reported many successes (Koppel et al., 2002; Koolen and van Cranenburgh, 2017; Qian, 2019), often driven by the fact that authors of specific genders tend to prefer producing content about topics that belie those gender (Sarawgi et al., 2011). Given

[3]We will interchangeably use the terms **speaker** and **author** here to refer to a creator of textual content throughout.

Dataset	M	F	N	U	Dim
Training Data					
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
Evaluation Data					
MDGENDER	384	401	-	-	ABOUT
MDGENDER	396	371	-	-	AS
MDGENDER	411	382	-	-	TO

Table 2: **Dataset Statistics**. Dataset statistics on the eight training datasets and new evaluation dataset, MDGENDERwith respect to each label.

this, we might additionally expect differences between genders along our *Speaking As* and *Speaking About* dimensions to interact, further motivating them as separate dimensions.

4 Creating Gender Classifiers

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 (Bolukbasi et al., 2016; Zhao et al., 2018b, 2019; Gonen and Goldberg, 2019)—and sometimes also explicitly (Caliskan et al., 2017; Hoyle et al., 2019)—gendered. While wordlist-based approaches provided a solid start on attacking the problem of gender bias, they are insufficient for multiple reasons. First, they conflate different conversational dimensions of gender bias, and are therefore unable to detect the subtle pragmatic differences that are of interest here. Further, all existing gendered word lists for English are limited, by construction, to explicitly binarily gendered words (e.g., *sister* vs. *brother*). Not only is binary gender wholly inadequate for the task, but restricting to explicitly gendered words is itself problematic, since we know that many words aren't explicitly gendered, but *are* strongly statistically gendered (see Table 1). Rather than solely relying on a brittle binary gender label from a global word list, our approach will also allow for gender bias to be determined flexibly over multiple words in context (Note: this will be crucial for examples that only receive gendered interpretations when in a particular context; for example, 'bag' disparagingly refers to an elderly woman, but only in the context of 'old', and 'cup' hints at masculine gender only in the context of 'wear').

Instead, we develop classifiers that can decompose gender bias over full sentences into semantic and/or pragmatic dimensions (*about/to/as*), additionally 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 for training these classifiers as well as details regarding the annotation of data for such training.

4.1 Models

We outline how these classifiers are trained to predict gender bias along the three dimensions, providing details of the classifier architectures as well as how the data labels are used. We train single-task and a multi-task classifiers for different purposes: the former will leverage gender information from each contextual dimension individually, and the latter should have broad applicability across all three.

Single Task Setting. 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).

Multitask Setting. To obtain a classifier capable of multi-tasking 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.

Model Architectures. For single task and multitask models, we use a pretrained Transformer (Vaswani et al., 2017) to find representations for the textual input and set of classes. Classes are scored—and then 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. The same architecture and pre-training as BERT (Devlin et al., 2018) are used throughout. We use `ParlAI` for model training (Miller et al., 2017). We will release data and models.

Model	About			To			As			
	Avg.	M	F	Avg.	M	F	Avg.	M	F	All Avg.
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 multi-task 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. MDGENDERwas collected to enable evaluation on the *masculine* and *femninine* 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 on the test sets from our training data.** We evaluate the multi-task model on the test sets for the training datasets. 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.

Model Labels. Many of our annotated datasets contain cases where the ABOUT, AS, TO labels are unknown. We retain these examples during training, but use two techniques to handle them. If the true label is *unknown* (for example, in Wikipedia, we do not know the gender of the author, so the *as* dimension is unknown), we either impute it or provide a label at random. For data for which the *about* label is unknown, we impute it using a classifier trained only on data for which this label is present. For data for which the *to* or *as* label is unknown, we provide a label at random, choosing between *masculine* and *feminine*. From epoch to epoch, we switch these arbitrarily assigned labels so that the model learns to assign the *masculine* and *feminine* labels with roughly equal probability to examples for which the gender is *unknown*. This label flipping allows us to retain greater quantities of data by preserving unknown samples. Additionally, we note that during training, we balance the data across the *masculine*, *feminine*, and *neutral* classes by oversampling from classes with fewer examples. We do this because much of the data is highly imbalanced: for example, over $> 80\%$ of examples from Wikipedia are labeled *masculine* (Table 2). We also early stop on the average accuracy across all three classes.

4.2 Data

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

Annotation of Existing Datasets. To enable training our classifiers, we leverage a variety of existing datasets. 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: inclusion of recoverable information about one or more of our dimensions, diversity in textual domain, and high quality, open data use. Once we narrowed our search to free, open source and freely available datasets, we maximized domain diversity by selecting datasets with high quality annotations along at least one of our dimensions (e.g., dialogue datasets have information on author and addressee gender, biographies have information on topic gender, and restaurant reviews have information on author gender).

The datasets are: Wikipedia, Funpedia (a less formal version of Wikipedia) (Miller et al., 2017), Wizard of Wikipedia (knowledge-based conversation) (Dinan et al., 2019d), Yelp Reviews[4], ConvAI2 (chit-chat dialogue) (Dinan et al., 2019c), ImageChat (chit-chat dialogue about an image) (Shuster et al., 2018), OpenSubtitles (dialogue from

[4] https://yelp.com/dataset

movies) (Lison and Tiedemann, 2016), and LIGHT (chit-chat fantasy dialogue) (Urbanek et al., 2019). We use data from multiple domains to represent different styles of text—from formal writing to chitchat—and different vocabularies. Further, several datasets are known to contain statistical imbalances and biases with regards to how people of different genders are described and represented, such as Wikipedia and LIGHT. Table 2 presents dataset statistics; the full detailed descriptions and more information on how labels were inferred or imputed in Appendix A.

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 is done can be found in Appendix A.

Collected Evaluation Dataset. We use a variety of datasets to train classifiers so they can be reliable on all dimensions across multiple domains. However, this weakly supervised data provides somewhat noisy training signal – particularly for the *masculine* and *feminine* classes – as the labels are automatically annotated or inferred. To enable reliable evaluation, we collect a specialized corpus, MDGENDER, which acts as a gold-labeled dataset.

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.[5] They are also provided with small sections of a biography from Wikipedia as the conversation topic. We observe that using biographies to frame the conversation encourages crowdworkers to discuss *about/to/as* gender information.

To maximize the *about/to/as* gender information contained in each utterance, we perform a second annotation over each utterance in the dataset. In this next phase, we ask 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 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, I went for a coffee with my friend and her dog* as the *her* indicates a woman. A rewrite such as *I went for a coffee with my friend* is not acceptable as it does not mention that the friend is a woman. After each rewritten utterance, evaluators label how confident they are that someone else would predict that the text is *spoken about, spoken as, or spoken to* a man or woman. For the rewritten utterance *I just got back from football practice*, many people would guess that the utterance was said by a man, as more men play football then women, but one cannot be certain (as women also play or coach football). An example instance of the task is shown in Table 9 and the interface is shown in Appendix Figure 2.

[5] We note that crowdworkers might perform genders in a non-authentic or idiosyncratic way when the persona gender doesn't match their gender. This would be an interesting avenue to explore in follow up work.

Model	Performance			
	M	F	N	Avg.
Multi-Task	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 masking out 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. To enable high quality evaluation along our proposed three dimensions, we use MDGENDERto evaluate. We measure the percentage accuracy for masculine, feminine, and neutral classes. We do not evaluate on the unknown class, as it is not modeled. 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 is another indication that gender information manifests differently along each dimension. Training for a single task allows models to specialize to detect and understand the nuances of text that indicates bias along one of the dimensions. However, in a multitask setting, models see additional data along the other dimensions and can possibly learn to generalize to understand what language characterizes bias across multiple dimensions.

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.

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 70% accuracy on predicting AS.

Analysis of Classifier Performance. We break down choices made during classifier training by comparing different models on the Wikipedia (ABOUT dimension). We train a single classifier of ABOUT, and train with the variations of masking out gendered words and names. As gendered words such as *her* and names are very 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. As expected, masking out gendered words and names makes it harder to classify the text, but the model is still able to obtain high accuracy.

6 Applications

We demonstrate the broad utility of our multi-task classifier by applying them 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 Wikipedia to find the most gendered biographies. Finally, we evaluate our classifier on an offensive text detection dataset to explore the interplay between offensive content and genderedness.

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, as well as for our baseline model trained using word lists. We measure the number of gendered words (from a word list) that appear in the generated text as well as the percentage of masculine-gendered words among all gendered words. Sequences are generated with top-k sampling, $k = 10$, with a beam size of 10 and 3-gram blocking.

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., 2017), generating questions in chit-chat (See et al., 2019), and reducing bias (Dinan et al., 2019a).

Previous work in gender bias used word lists to control bias, but found that word lists were limited in coverage and applicability to a variety of domains (Dinan et al., 2019a). 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 multi-task 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 gender label along the dimension. For example for the ABOUT dimension, we prepend utterances with tokens *ABOUT:<gender_label>*, where *<gender_label>* denotes the label assigned to the utterance via the classifier. 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. (2019a), we use 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 Reddit. At inference time, we generate text via top-k sampling (Fan et al., 2018), 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 control tokens (as well as the word list baseline) are shown in Table 10 in the Appendix. These examples illustrate how controlling for gender over different dimensions yields extremely varied responses, and why limiting control to word lists may not be enough to capture these different aspects of gender. For example, adjusting AS to 'feminine' causes the model to write text such as *Awwww, 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 more 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 likely due to the distribution of such gendered words in the training data for the classifier in the *to* and *as* dimensions. 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 drops to one-third.[6] By design, the word list baseline has the best control over whether the generations contain words from this word list. These results, as well as the previously described qualitative results, demonstrate why evaluating and controlling with word lists is insufficient — word lists do not capture all aspects of gender.

Percentage of masculine-gendered text				
Dim	*Safe*	*Offensive*	*t-statistic*	*p-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 7: **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.

6.2 Bias Detection

Creating classifiers along different dimensions can be used to detect gender bias in any form of text, beyond dialogue itself. We investigate using the trained classifiers to detect the most gendered sentences and paragraphs in various documents, and analyze what portions of the text drive the classification decision. Such methods could be very useful in practical applications such as detecting, removing, and rewriting biased writing.

Methods. We apply our classification models by detecting the most gendered biographies in Wikipedia. We use the multitask model to score each paragraph among a set of $65,000$ Wikipedia

[6]The 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)

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

Table 8: **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.

biographies, where the score represents the probability that the paragraph is *masculine* in the *about* dimension. We calculate a *masculine genderedness score* for the page by taking the median among all paragraphs in the page.

Quantitative Results. We report the average and median *masculine genderedness scores* for all biographies in the set of $65,000$ that fit this criteria, and for biographies of men and women in Table 8. 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—containing a minimum of 25 paragraphs—with the minimum score (most feminine-gendered biographies) and the maximum score (most masculine-gendered biographies) in Table 11 in the Appendix. We observe that the most masculine-gendered biographies are mostly composers and conductors, likely due to the historical gender imbalance in these occupations. Amongst the most feminine gendered biographies, there are many popular actresses from the mid-20th century. By examining the *most* gendered paragraph in these biographies, anecdotally we find these are often the paragraphs describing the subject's life after retirement. For example, the most gendered paragraph in 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 interesting area for study, as many examples of gendered text—be they explicitly or contextually gendered—are disparaging or have negative connotations (e.g., "cat fight" and "doll"). 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 offensive language in the context of dialogue (Dinan et al., 2019b). We investigate this relationship by examining the distribution of labels output by our gender classifier on data that is labeled for offensiveness.

Methods. For this application, we use the *Standard* training and evaluation dataset created and described in Dinan et al. (2019b). We examine the relationship between genderedness and offensive utterances by labeling the gender of utterances (along the three dimensions) in both the "safe" and "offensive" classes in this dataset 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 7. We observe that, on the *self* and *partner* 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 the masculine 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 in text by decomposing it along three dimensions: (1) gender of the person or people being spoken about (ABOUT), (2) gender of the addressee (TO), and (2) gender of the speaker (AS). We show that classifiers can detect bias along each of these dimensions. We annotate eight large existing datasets along our dimensions, and also contribute a high quality evaluation dataset for this task. We demonstrate the broad utility of our classifiers by showing strong performance on controlling bias in generated dialogue, detecting genderedness in text such as Wikipedia, and highlighting gender differences in offensive text classification.

References

Fran Amery, Stephen Bates, Laura Jenkins, and Heather Savigny. 2015. Metaphors on women in academia: A review of the literature, 2004-2013. *At the center: Feminism, social science and knowledge*, 20:247À267.

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

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.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2020. *Fairness in machine learning: Limitations and Opportunities*.

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*.

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.

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.

Kay Bussey. 1986. The first socialization. In *Australian women: New feminist perspectives*, pages 90–104. Oxford University Press.

Judith Butler. 1990. *Gender trouble, feminist theory, and psychoanalytic discourse*. Routledge New York.

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.

Deborah Cameron. 1990. The feminist critique of language: A reader.

Yang Trista Cao and Hal Daumé. 2019. Toward gender-inclusive coreference resolution. *arXiv preprint arXiv:1910.13913*.

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.

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.

Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. 2011. Author gender identification from text. *Digital Investigation*, 8(1):78–88.

Jennifer Coates. 2015. *Women, men and language: A sociolinguistic account of gender differences in language*. Routledge.

Marta R Costa-jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, pages 1–2.

Mary Crawford. 1995. *Talking difference: On gender and language*. Sage.

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.

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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019a. Queens are powerful too: Mitigating gender bias in dialogue generation.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019b. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. *arXiv preprint arXiv:1908.06083*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019c. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019d. 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 6135–6145.

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 Sally McConnell-Ginet. 2013. *Language and gender*. Cambridge University Press.

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.

Angela Fan, David Grangier, and Michael Auli. 2017. Controllable abstractive summarization. *arXiv preprint arXiv:1711.05217*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Almudena Fernandez Fontecha and Rosa Maria Jimenez Catalan. 2003. Semantic derogation in animal metaphor: a contrastive-cognitive analysis of two male/female examples in english and spanish. *Journal of pragmatics*, 35(5):771–797.

Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas. 2006. Effective identification of source code authors using byte-level information. In *Proceedings of the 28th international conference on Software engineering*, pages 893–896.

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. Womens syntactic resilience and mens 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.

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, et al. 2019. Towards understanding gender bias in relation extraction. *arXiv preprint arXiv:1911.03642*.

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 Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? *arXiv preprint arXiv:1910.14161*.

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.

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.

Janet Holmes. 2013. *An introduction to sociolinguistics*. Routledge.

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. Can you translate that into man? commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

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. *arXiv preprint arXiv:1906.00742*.

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (male, bachelor) and (female, Ph.D) have different connotations: Parallelly 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.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412.

George Lakoff and Mark Johnson. 1980. Metaphors we live by. *Chicago, IL: University of Chicago*.

Robin Lakoff. 1973. Language and woman's place. *Language in society*, 2(1):45–79.

Robin Lakoff. 1990. *Talking Power: The Politics of Language*.

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.

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.

Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 513–520.

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. *CoRR*, abs/1909.00871.

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.

Thomas Corwin Mendenhall. 1887. The characteristic curves of composition. *Science*, 9(214):237–249.

Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Sara Mills. 2014. *Language and gender: Interdisciplinary perspectives*. Routledge.

John Money and Anke A Ehrhardt. 1972. Man and woman, boy and girl: Differentiation and dimorphism of gender identity from conception to maturity.

Rosamund Moon. 2014. From gorgeous to grumpy: adjectives, age and gender. *Gender & Language*, 8(1).

Frederick Mosteller and David L Wallace. 1984. *Applied Bayesian and classical inference: the case of the Federalist papers*. Springer Verlag.

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.

Jian Peng, Kim-Kwang Raymond Choo, and Helen Ashman. 2016. User profiling in intrusion detection: A review. *Journal of Network and Computer Applications*, 72:14–27.

Yusu Qian. 2019. Gender stereotypes differ between male and female writings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 48–53.

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. *arXiv preprint arXiv:1905.12801*.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 38–42, Uppsala, Sweden. 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.

Anderson Rocha, Walter J Scheirer, Christopher W Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne RB Carvalho, and Efstathios Stamatatos. 2016. Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, 12(1):5–33.

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, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019a. 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.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019b. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86, Portland, Oregon, USA. 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. *arXiv preprint arXiv:1902.08654*.

Sima Sharifirad, Alon Jacovi, Israel Bar Ilan Univesity, and Stan Matwin. 2019. Learning and understanding different categories of sexism using convolutional neural networks 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*.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 1999. Automatic authorship attribution. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway. Association for Computational Linguistics.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos. 2017. Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1138–1149, Valencia, Spain. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

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.

Jane Sunderland. 2006. *Language and gender: An advanced resource book*. Routledge.

Joan Swann. 1992. *Girls, boys, and language*. Blackwell Publishers.

Mary Talbot. 2019. *Language and gender*. John Wiley & Sons.

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.

Ann Weatherall. 2002. *Gender, language and discourse*. Psychology Press.

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 Existing Data Annotation

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.[7] 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.
3. **Wizard of Wikipedia** - Wizard of Wikipedia (Dinan et al., 2019d) 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 content of an image. We use the (Xu et al., 2015) image captioning system[8] 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 author 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., 2019c) 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. **OpenSubtitiles** - OpenSubtitles[9] (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 States[10]. 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 labeled the gender of each persona (Dinan et al., 2019a), 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.

B New Evaluation Dataset

The interface for our new evaluation dataset MD-GENDER can be seen in Figure 2. Examples from the new dataset can be found in Table 9.

C Applications

Example generations for various control tokens, as well as for our word list baseline, are shown in Table 10. See §6.1 on Controllable Generation in the main paper for more details.

The top 10 most gendered Wikipedia biographies are shown in Table 11. See §6.2 on Detecting Bias in the main paper for more details.

[7] This 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.

[8] https://github.com/AaronCCWong/Show-Attend-and-Tell

[9] http://www.opensubtitles.org/

[10] https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data

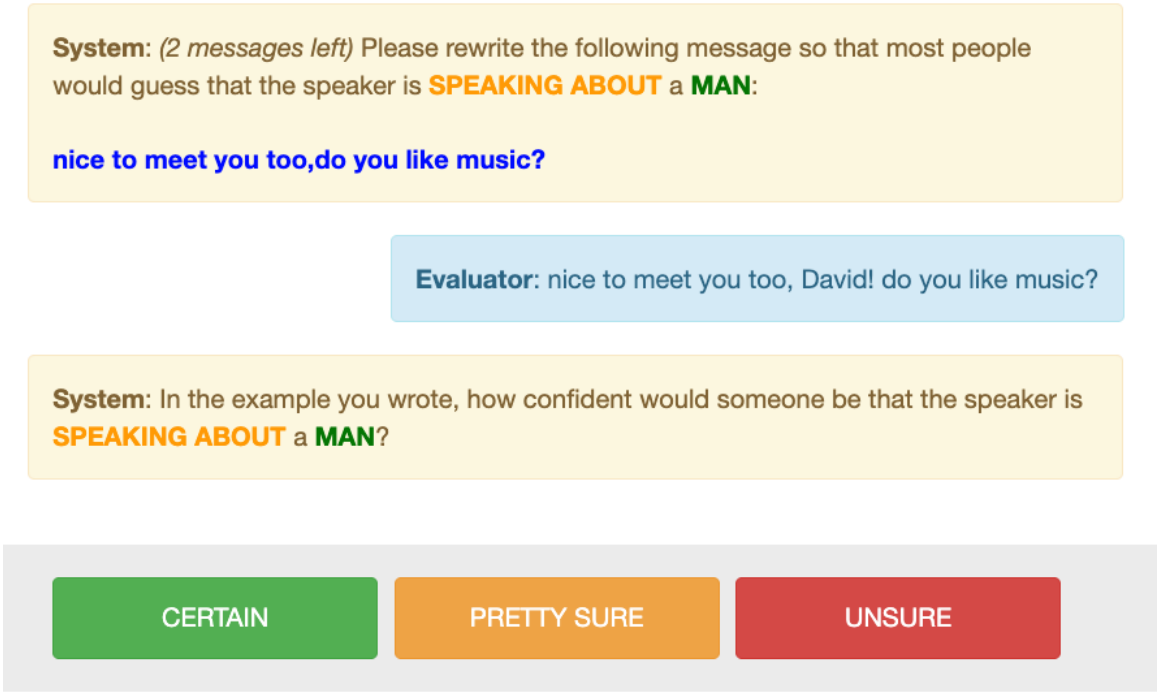

Figure 2: **Annotation interface**. Annotation interface for collecting MDGENDER. Annotators were shown an utterance from a conversation, and asked to re-write it such that it is clear they would be speaker about/to/as a man or a woman. They were then asked for their confidence level.

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

Table 9: **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).

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** Awwww, 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
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.
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.
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:male** 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.
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.
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.

Table 10: **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.

Most Feminine	*Most Masculine*
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 Martin:** 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. **Chlo 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 11: **Most gendered Wikipedia biographies** We ran our multi-task 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.

.