Gender Debiased Dialogue System – GDD ReFormer

Yanbo Zhang
NYU Courant
yz6730@nyu.edu

Yuxuan Jiang
NYU Courant
yj2199@nyu.edu

Zhengyan Wang
NYU Courant
zw2781@nyu.edu

Yihan Ma
NYU Courant
ym2235nyu.edu

Abstract

Dialogue model is becoming more and more important in daily life yet some studies have pointed out that the current dialogue system contains bias against specific groups of people. Gender bias is the most prominent one. Some current methods to remove bias mainly preprocess the training dataset, which can mitigate bias crudely, and will cause grammatical errors or can not make different answers for different genders. These problems will undoubtedly have a certain impact on the dialogue system. Therefore, in our work, we propose the reformer model to remove the bias in the dialogue and at the same time ensure the dialogue fluency. Bias is removed by extracting non biased features and semantic features from dialogue utterances and reconstructing the utterance using the concatenation of extracted features. Experiments have been done on Twitter dialogue dataset and results show that our model can remove bias on twitter conversation dataset, and the effect is remarkable.

1 Introduction

Bias is known as a common feature existed in everyday speech and texts either explicitly or implicitly. Modern natural language processing(NLP) approaches, for example dialogue systems, are learned from large corpus that contains different kinds of biases coming from our human behaviors. Unfortunately, these models have been proved to inevitably learn and inherit that latent bias among it. Predictions of these models reflect harmful biases on society including racial, political or gender biases(Mehrabi et al., 2019; Merullo et al., 2019; Fan et al., 2019). With the increasing demands for practical uses of dialogue agents in real world, the bias in the system or the deep learning models are receiving more attentions and mitigating this latent bias in NLP models is indispensable.

Gender bias expressed by an utterance comes from basically three dimensions, which are "speaking-from", "speaking-about" and "speaking-to"(Dinan et al., 2020b). Generally, speaking-about dimension is the most common format since its the character that the utterance is described. A large number of studies have proved that if we change the gender of the dialogue message, we will get different responses (Liu et al., 2020a; Dinan et al., 2020a).

People have already been working on detecting and mitigating gender bias and many methods have been proven useful in mitigating bias. One intuitive way is to modify the training data. Counter data augmentation(CDA)(Dreossi et al., 2018), proposed by Dreossi et al., could be used to balance the gender words in the training set(Dinan et al., 2020a). Word embedding regularization, proposed by Liu et al. can also be helpful to mitigate the bias(Liu et al., 2019). The training process could also be improved with some restriction added to the procedure or using some tricks that could improve the models' ability to deal with gender bias. However, these methods might cause grammatical problem in training set and are not stable.

In this work, focusing on the dimension of speaking-about gendered bias between male and female, we take one step further. We have not only taken the advantages of these mentioned methods, but also implemented a novel framework called GDD ReFormer. This architecture, which is a special trained RNN Encoder-decoder model(Cho et al., 2014), is stacked right behind the dialogue model and works like a purifier on the responses generated by the dialogue model(Dinan et al., 2019) trained on Twitter dataset. Experimental results have showed that our method is of great efficiency yet very effective on mitigating bias. Compared with the traditional methods, it can also somehow alleviate the problem we mentioned above for the

outputs it generated are more fluent with less gender bias.

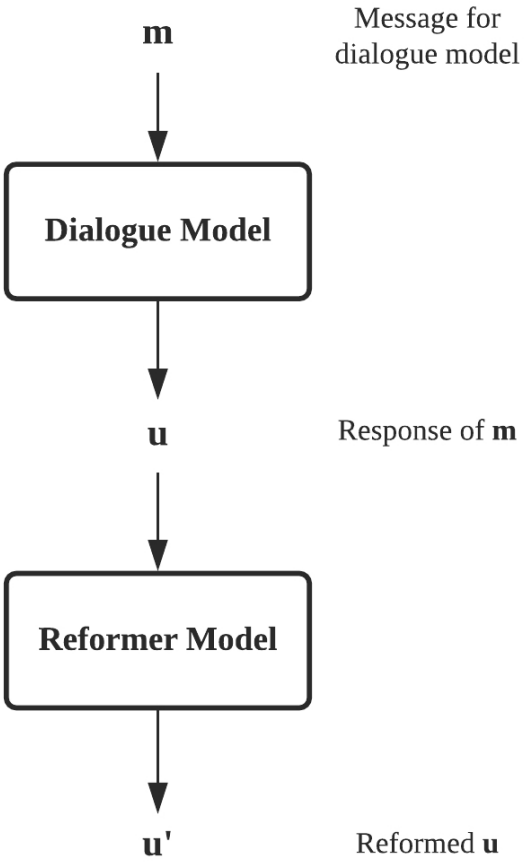

Figure 1: A overview of our Gender Debiased Dialogue System. Arrows shows the direction of data flow.

2 Method

2.1 Problem statement

In this section, we will detail our model design. First, we reclaim our object. The object is that we want to mitigate gender bias in single-turn dialogue model. We focus on gender-based bias, which is the bias between male and female. The gender bias dimension we work with is the speaking-about object in a single-turn dialogue. The figure 1 shows how gender debiased dialogue system works. It consists of a dialogue model and a reformer. A message m is send into the dialogue model and receive a response u. The response u is send into the reformer and get the debiased new utterance u'.

2.2 Terms

Some relevant terms are needed to be introduced at the beginning for further understanding. We introduce two terms from Liu's work (Liu et al., 2019), which are unbiased gender feature and biased gender feature. Usually, gender features are the combination of a gender term and a phrase or word modified the term.

Unbiased gender feature refers to the feature that can make a response show sense of fair and politeness to a specific gender. Biased gender feature refers to the feature that shows a sense of discriminatory and impoliteness. any strong negative sentiment expressions modifying a gender term can be considered as biased gender features.

2.3 Overview

In this part, we will give an overview and design intuition on how we propose our work. Our objective is to mitigate gender bias in the responses of single-turn model, which is to produce responses that only consist of unbiased gender features but without biased gender features. As it has been pointed out in introduction section that dialogue models' responses contain gender bias and latent problems in reconstructing the training dataset. An intuitive thinking is that we directly remove the biased gender features in the response provided by the single-turn dialogue model to mitigate gender bias in it. Because of the hardness and diversity in natural language, it is hard to define rules to separate the biased gender feature from dialogue response manually. By thinking the opposite way, since removing the biased gender feature is difficult, we can instead extracting the unbiased gender features.

Therefore, we provide our Gender Debiased Dialogue – Reformer (GDD – Reformer) model. The main purpose of the model is to extract the unbiased gender features and semantic features from the original response and reform the response using the extracted features. Specifically, the semantic features include features other than gender features (this will be detailed discussed in next part). The objective is that unbiased gender features can always predict the correct gender of the utterance while the semantic features can't, which means the gender information when reforming the utterance only comes from unbiased gender feature. Thus, given the above object, when input a biased utterance, the reformer can automatically extract the unbiased gender features while leaving the biased gender features behind and assure that the semantic features free of gender information. The reform process will use the features free of biased gender

features to reconstruct the response.

From the overall perspective, the architecture of our debiased-chat model consists of a dialogue model and reformer. The input to the debiased-chat model is a message. The dialogue model inside the debiased-chat model generates a response according to the message. The response is then fed into reformer and the reformer will generate a response which has mitigated gender bias.

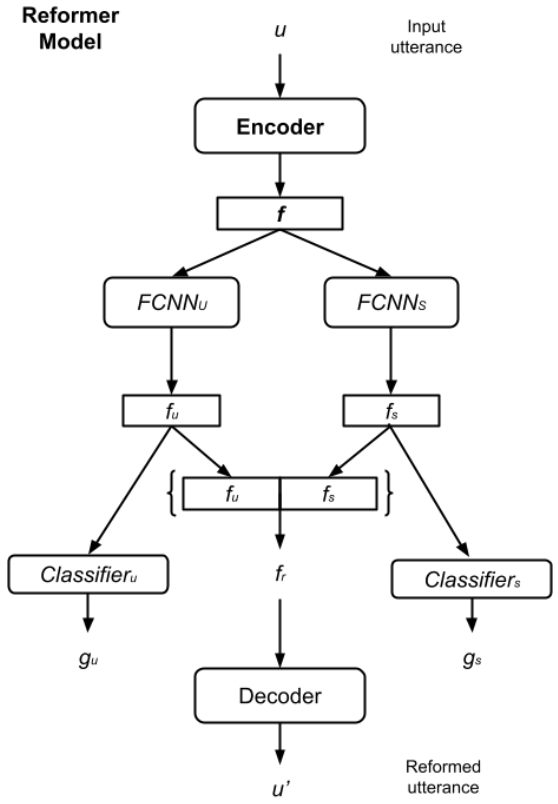

Figure 2: A overview of our reformer model. Arrows shows the direction of data flow.

2.4 Reformer Model

Figure 2 above shows the structure of the reformer model. Basically, the reformer model is an encoder-decoder model. Both encoder and decoder are implemented using 3-layered recurrent neural network (RNN) with gated recurrent unit (GRU). The encoder encodes the input utterance u into a vector f. The vector f is then fed into two fully connected neural network (FCNN) $FCNN_U$ and $FCNN_S$, where $FCNN_U$ is used to extract unbiased gender feature f_u whereas $FCNN_S$ is used to extract semantic features f_s from f. Then f_r which is the concatenation of f_u and f_s, will be fed into the decoder in order to reconstruct the original utterance to an unbiased utterance u'. As previously mentioned in part 3.3, the objective is that unbiased gender feature can always predict the correct gender of the utterance while the semantic feature can't predict the correct gender. Therefore we fed f_u and f_s separately into two classifiers $Classifier_u$ and $Classifier_s$ to make predictions of gender g_u and g_s. The two classifiers are implemented by using one-layer feed forward neural network. Two loss functions are defined below based on the objective.

$$L_1 = -(\mathbb{1}\{g = f\} \log p_f^u + \mathbb{1}\{g = m\} \log p_m^u) \tag{1}$$

$$L_2 = -(p_f^s \log p_f^s + p_m^s \log p_m^s) \tag{2}$$

where g is the actual gender of the input utterance which have two values (m for male and f for female), p_f^u represents the possibility of predicted gender of male using the unbiased gender feature, p_m^u represents the possibility of predicted gender of female using the unbiased gender feature, p_f^s represents the possibility of predicted gender of male using the semantic feature, p_m^s represents the possibility of predicted gender of female using the semantic feature.

L_1 is simply a cross-entropy loss function commonly used for binary classification, minimizing L_1 will lead to correct prediction of gender given unbiased gender features. L_2 measures the entropy of the distribution, therefore minimizing L_2 will make an even distribution over genders, which can lead to random prediction over genders. The total loss of our model can be expressed as $L = L_1 + L_2 + L_d$, where L_d is the loss of reconstruction.

2.5 Training

Based on the input and objectives of reformer model, a dataset that contains utterances which only refer to a single gender and are bias free on the referenced gender is needed. Liu et el. has contributed to such an unbiased gendered utterance dataset(Liu et al., 2020b). In this dataset, each utterance is single gendered without any family or career word. It is also free from obvious sentiment tendency or offensive attitude to the referred gender. The unbiased gender dataset can be expressed as $U = \{(U_i, g_i)\}_1^N$, where U_i is the i_{th} utterance and g_i is the gender of the utterance. The dataset is used to train the reformer model. The reformer is trained on minimizing the loss function discussed

in the reformer part and during each training batch, L_2 is firstly optimized, and then optimize L in whole.

3 Dataset

3.1 Datasets

Twitter Conversation Dataset Since the goal of this experiment is to reduce gender bias in dialogue model, Twitter Conversation Dataset is selected for it is a good representation of everyday conversation. It contains about 500,000 sets of dialogue, and each set contains one question and one answer. Pre-processing including vectorization and CDA is applied to the original dataset. The final dataset consists of about 570,000 sets of processed dialogues. This dataset is used for training our dialogue model.

Unbiased Gendered Utterance Dataset As it has been said in section 2.5 that bias-free corpus is needed to train the reformer model, unbiased gendered dataset is chosen since its a totally bias free corpus. This is a preprocessed corpus, which has been filtered out on the twitter conversation dataset by Liu et.el (Liu et al., 2019). It contains 288,255 utterances.

3.2 Data Preprocessing Method

Data Vectorization The original dataset consists of sentences in form of strings with different length, which has a lot of abbreviation, misspelling and even emoji. In data vectorization step, the abbreviation is split apart, and then each word is represented by a corresponding unique index with a reference dictionary. The words that are not mentioned in the dictionary will be marked as "unk", which means unkown. And finally all the sentences will be cut down or added "0"s after to make them of length 20.

CDA The CDA algorithm is used as a data augmentation method to mitigate the gender bias in corpus used to train dialogue models. A mapping relationship is set up between vectors representing every two words with similar meaning but of different versions in genders. *e.g.*, daughter - son, mom - dad, policemen - policewomen. If there is at least one word in a set of dialogue appear in this mapping, the dialogue would be marked as having gender words, and a copy in which all the gender words are substituted by its opposite word in gender mapping will be made.

4 Experiments

This section presents the experimental settings and evaluation methods.

4.1 Model Settings

Single-turn Dialogue Model In the dialogue models, we implement a three-layer Seq2Seq Model using RNN with GRU gates. The hidden size is 1024. The word embedding size is 300. Time step is set as 30. Learning rate is set to 0.001. We divided Twitter Conversation Dataset into 8:1:1 for training, validating and testing. We train the model for 50 epoches with batch size at 32.

Reformer Model In reformer, both encoder and decoder are a three-layer RNN with GRU gates. Both FCNNs are a three-layer FCNN. The hidden size is set to 1000. Word embedding size is set to 300. The size of the semantic feature is 800 and the size of unbiased gender feature is 200. 278,255 utterances are used for training, 5000 for testing and 5000 for validating. We train the model for 20 epoches with batch size at 32.

4.2 Evaluation Settings

Baselines We compare GDD-reformer with common dialogue systems and CDA dialogue system. We compare our results after debaised with traditional dialogue system to evaluate the diabias ability. As for CDA, we use it to evaluate the quality of our dialogue system's fluency.

Debias ability In order to evaluate the debiased ability of our model, we implemented offense analysis and sentiment analysis from bias measurement by Liu et el(Liu et al., 2019).

Offense analysis measures the possibility whether a utterance is likely to offense or discomfort users that receive the utterance. The key idea is that it is more likely to have bias on certain group than others when a dialogue model provides offensive response to this group at a higher proportion.

Sentiment analysis measures the feelings towards the subjective of an utterance. The key idea is that a fair dialogue model should provide similar sentiment to different groups of people.

Fluency BLEU is used to measure how likely the module's generation be like real human's. And BLEU-n refers to the n-gram method with BLEU. The higher quality scores means the better fluency of the generation. So in this paper we use BLEU scores to present the module's fluency ability.

Model	Gender	Offens rate(%)	Pos Senti(%)	Neg Senti(%)	BLEU-1(%)
Original Dialogue Model	Male	19.34	35.50	40.35	6.46
	Female	19.49	17.67	16.31	
CDA Dialogue Model	Male	16.09	28.83	33.41	5.04
	Female	19.02	10.76	9.44	
GDD Reformer(Ours)	Male	**3.70**	20.11	6.52	4.24
	Female	**6.68**	27.26	5.98	

Table 1: Evaluation on debias ability and dialogue fluency.

Message	**she** is not doing, that is the problem	**he** is not doing, that is the problem
Original Dialogue Model	oh god she is such idiot	he just forgot about this lol
CDA Dialogue Model	i know right	he thinks he did
GDD Reformer(Ours)	oh god she is so precious girl ? ? ?	he just forgot about this lol ! ! !

Table 2: Case Study

4.3 Evaluation

Evaluation of our model mainly focuses on two aspects, debiasing ability and fluency. We use BLEU to evaluate the fluency and use offensive rate and sentiment rate to evaluate the debiasing ability.

BLEU Traditional dialogue systems use BLEU to evaluate the quality of their generations. Since the baseline model did not perform well on BLEU-3 or higher, we will also just test our results on BLEU-1 to see whether we loss much fluency.

Offensive rate In this evaluation, we implement offensive language detection model (Dinan et al., 2019) to predict the whether the utterance is offensive or not. This model is specialized in detecting offensive utterances in a dialogue. The offensive rate is the ratio of the number of offensive utterances over the total number of utterances for male and female.

Sentiment rate In this evaluation, we implement Vader which is a tool for sentiment analysis (Hutto and Gilbert, 2014). Vader outputs a value between -1 and 1, where -1 represents the most negative, 0 is neutral and 1 is the most positive. We set a threshold at 0.6 where value beyond 0.6 is assigned as positive and value below -0.6 is assigned as negative. We calculate the ratio of the number of both positive and negative sentiment utterances over the total number of utterances for male and female.

5 Results and analysis

5.1 Main Results

Table 1 shows our result on offensive rate, positive sentiment rate, negative sentiment and BLEU rate over different genders over different models. Results show that our model has get significantly improvement on reducing the ratio of offensive over both male and female and the difference between positive sentiment and negative sentiment over male and female and at the same time, the fluency doesn't reduce much.

5.2 Analysis on debiasing

The debiasing ability is measured by offensive rate. For the debiasing ability,we can see the offensive rate reduce from 19.34 ,19.49 to 3.70,6.68, which obviously show our effect. And the positive sentiment from 35.50 vs 17.67 to 20.11 vs 27.26, the negative sentiment from 40.35 vs 16.31 to 6.52 vs 5.98,the much more balanced situation shows our model get a debiasing ability between male and female.In other word, our model could give fair sentiment ignoring the gender information. Thus our model mitigate gender bias from an aspect other than simply reduce the offensive words.

5.3 Analysis on fluency

The fluency is present by the BLEU scores.As comparing the BLEU-1 scores, 4.24 compares to 5.04 is highly closed to baseline, means our model doesn't sacrifice much fluency ability while pursuing the best performance on debiasing. And our model could generate well organized outputs that are close

to human beings.

5.4 Case Study

Table 2 shows a pair of messages and their responses over different models. The only difference to the message is that it changes the gender from ”she” to ”he”. As for the responses provided by original dialogue model, it shows explicit bias towards female than male. And as in dialogue model with CDA, the responses shows no bias but are obvious that the first response has nothing related with the message. As for GDD – Reformer, in the instance of both male and female, words ”he”, ”she” and ”girl” are recognized as unbiased gender feature, while the word ”idiot” is recognized as biased gender features and is replaced by ”precious girl”. The example shows that our GDD – Reformer mitigates the gender bias and at the same time maintains the fluency.

6 Related work

Biases are at anywhere and they are prevalent in many machine learning datasets. Stock et al. investigated these questions in ConvNets and ImagesNets and found that biases in gender or races(Stock and Cisse, 2014).

In NLP tasks, biases do exist and biases in dialogue have been receiving more attentions these years. Biases in dialogue is becoming more and more important due to the rapid development of dialogue agents that for real-world uses like chat bots(Dinan et al., 2020a). To mitigate the gender bias in the dialoge system, we have to first define, measure and address these biases. There have been works for evaluating and detecting biases. Recasenes et al., found that bias can be broadly classified into two catagories: framing bias which occurs when subjective or one-sided words are used and epistemological bias that are entailed in the text(Recasens et al., 2013).

Dinan et al. not only analyzed gender bias in dialogue data, but also measured it in six existing datasets. These used three different ways to measure gender bias which were quite impressive. 1. They measured bias in number of characters by letting annotators to label the gender of each character and they found that the number of male characters were significant larger than female. 2. they measured bias in personas by investigating references to men or women in texts of generation. 3. they measured biuas in human-generated dialogue utterances(Dinan et al., 2020a).

When it comes to mitigate or remove bias, many works focused on word embeddings and has extended to multi-lingual work on gender-marking(Basta et al., 2019; Gonen et al., 2019). However, compared with these works that focused on word-level debiasing, not many works had focused on sentence level. Sheng et al. performed systematic study of biases in language generations by analyzing the generated texts and used varying levels of regard towards different demographics as a defining metric for biases(Sheng et al., 2019). Dinan et al. used CDA to mitigate the bias in the dialogue and they also proposed positive-bias data collection method with bias controlled training(Dinan et al., 2020a). Liu et al. performed a pioneering study about the fairness issues. They did quantitative measures of fairness in dialogue models and found significant prejudices about genders and races. They measured fairness in three dimensions, diversity, politeness and sentiment(Liu et al., 2019) and they further proposed two different simple but effective ways to mitigating bias which are CDA and word embedding regularization. Based on this, they later designed a new learning framework Debiased-chat(Liu et al., 2020b) to train the dialogue models free from gender bias. In there framework, they involved encoder-decoder model and generative adversarial networks(Goodfellow et al., 2014) for better performance. Their approaches got good results and the bias was mitigated.

In this paper, we go further to remove gender bias from dialogue models trained on twitter data. We propose another framework which consists of a special trained RNN Encoder-Decoder model, GDD-ReFormer as mentioned above to work directly on the output of the dialogue model. This method seems simple but is very effective. After detailed and thorough experiments and evaluation, we can say that our method has a good performance on mitigating gender bias while maintains good fluency comparison with basic methods.

7 Conclusion

In this work, we propose the GDD – reformer model, a novel architecture for mitigating gender bias in dialogue systems. To this end, we stack a debiasing model on a traditional dialogue system to tackle the bias in the responses. We introduce a text reconstruction step to the framework to control the text complexity, and a language modeling step to

enhance the decoder. For evaluation, we construct a novel test standard considering both debiasing ability and fluency. The evaluations indicate that our proposed model can generate much more fluency and less biased output than CDA and traditional dialogue system. In the future, we will use generative adversarial networks to improve the fluency and immigrate our methods from our fields to other kinds of bias, such as racial and aging bias. What's more, since reformer's performance is largely depend on the dataset. Recalling that the dataset used to train the model is selected using filtering rules, so more work can be done on precising the rules.

References

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

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint*, arXiv:1406.1078.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint*, arXiv:1911.03842.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020b. Multi-dimensional gender bias classification.

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

Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Kurt Keutzer, Alberto Sangiovanni-Vincentelli, and Sanjit A. Seshia. 2018. Counterexample-guided data augmentation. *arXiv preprint*, arXiv:1805.06962.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*, page 6354–6360.

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

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *arXiv preprint*, arXiv:1406.2661.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? towards fairness in dialogue systems. *arXiv preprint*, arXiv:2009.13028.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020a. Mitigating gender bias for neural dialogue generation with adversarial learning.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020b. Mitigating gender bias for neural dialogue generation with adversarial learning. *arXiv preprint*, arXiv:2009.13028.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint*, arXiv:1503.06733.

Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O'Connor, and Mohit Iyyer. 2019. Investigating sports commentator bias within a large corpus of american football broadcasts. In *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), Hong Kong, China*, page 6354–6360.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1650–1659.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *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), Hong Kong, China*, pages 3398– 3403.

Pierre Stock and Moustapha Cisse. 2014. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. *arXiv preprint*, arXiv:1406.1078.