Improving Performance of Multi-Dimensional Gender Bias Classification

Aishwarya Raman
Department of Computer Science
New York University
New York, NY 10012
ar6381@nyu.edu

Hyejin KIM
Department of Computer Science
New York University
New York, NY 10012
hk3342@nyu.edu

Yuan Shen
Department of Computer Science
New York University
New York, NY 10012
ys4420@nyu.edu

Abstract

In statistical NLP, a trained model is used to find patterns in the data. Gender bias in data could also be treated as pattern and learnt by the model. In this study we categorize bias in NLP tasks into three dimensions: bias of gender the person is being spoken about, bias of gender the person is speaking as and bias of gender the person is speaking to. We classify the dataset with these three dimensions and label them as male or female. We perform three regularization techniques, variant of label smoothing, data augmentation, mix-out augmentation, based on our hypothesis to improve the model performance. To analyze the impact of each regularization to the model performance, we compare the accuracy of each method. We further discuss how to mitigate data bias based on this work.

1 Introduction

Language is a way people communicate with each other. People use language to identify themselves and others. During that case people may use words to explicitly categorize the genders of those they are speaking to like pronouns such as "her", nouns such as "businessman". People would also express genders of themselves in a similar manner. However there are also words with which people might implicitly address the gender. Some non-arguable examples including king and queen, as a king is masculine and a queen is feminine. Other examples, as shown by Bolukbasi et al.[1], might indicate a gender bias in the words we use. In that paper they show that (computer programmer, homemaker) is almost analogous to (man, woman) as (king, queen) is analogous to (man, woman). We could not and should not imply the gender of a person given words like computer programmer and homemaker. However in real-world examples, such connections might be indicated by the language we use. A gender bias in our language is thus introduced. A NLP model trained on a biased data set can learn the bias and even amplify that bias during prediction. The consequence of such bias would be reinforcing stereotypes with applications of those models.

Dinan et al.[2]. try to classify the data set with three dimensions,

1.1 Multi dimensional classification

1.1.1 Speaking about

Speaking about category refers to the gender of the person that is being talked about. People might choose different adjectives and verbs depending on the gender of the person they are describing.

1.1.2 Speaking as

Speaking as class refers to the gender of the speaker himself/herself. Gender is an important category of self-identification and how others might interact with them. These will affect the word they choose, how they address a sentence, etc.

1.1.3 Speaking to

Speaking to class refers to the gender of the addressee of the speaker. The gender of the listener might also influence the word and the topic they choose.

However, even though the multi-dimensional classification result on test data from Dinan et al.[2] shows decent accuracy for Wikipedia(avg 77.22) and Funpedia(avg 76.2), it doesn't explain the relatively low accuracies for other datasets including ConvAI2(avg 53.63), OpenSubtitles(avg 57.01). Also, it doesn't contain further works to improve the model. In this study, we experiment several regularization techniques to improve the model performance of multi-dimensional gender bias classificatio task. We suggest three hypothesis to improve the model performance and feasible strategies for each hypothesis.

1.1.4 Obervation on dataset sizes and results

Comparing the datasets which showed relatively lower performance– The LIGHT(Urbanek et al., 2019)[8] with avg 58.65 accuracy, ConvAI2(Dinan et al., 2019c)[4] with avg 53.63 accuracy, and ImageChat(Shuster et al., 2018)[6] with avg 51.09 accuarcy– to the datasets such as Wikipedia with higher accuracy(avg 77.22 accuracy), there was considerable difference in their dataset sizes; LIGHT has total 104K data with only 21K data for Male/Female labels(i.e. 83K for Unknown), ConvAI2 is total 130K with only 44K for Male/Female labels, and ImageChat is size 208K with only 54K for Male/Female labels, while the size of Wikipedia is 12M. From these observations, we established the hypothesis that augmenting data would help improve the model accuracy.

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

Figure 1: Performance

1.1.5 Observation on labels

There are 6 possible labels in multi-dimensional gender bias classification; male, femal x about, as, to. For each dimension(about, as, to), labels are more similar to the same gender bias(i.e. 'About:Female' is more similar to 'As:Female' than 'About:Male'). By reflecting this similarity to smooth the labels, we may be able to decrease model uncertainty and further improve the model performance.

1.1.6 Using pretrained bert model

We train our classifier for the multi-dimensional gender bias classification using large-scale pre-trained DistilBert models. Another approach to get better regularized model would be fine-tuning DistilBert model using our train dataset. And our train dataset size would be relatively smaller than the large corpus the model is pretrained on. Mixout, a novel regularization strategy for fine-tuning BERT suggested by Lee et al. [13] shows that fine-tuning BERT with mixout improves accuracy and overall performance of the model especially when there are only a small number of training instances available. So we established another hypothesis that fine-tuning DistilBert with mixout could improve model performance for our task.

2 Model

BERT stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. [14]

DistilBERT is a small, fast, cheap and light Transformer model trained by distilling BERT base. It has 40% less parameters than bert-base-uncased, runs 60% faster while preserving over 95% of BERT's performances as measured on the GLUE language understanding benchmark. [15]

In this project we developed our own model for classification based on pre-trained DistilBert.

3 Dataset

3.1 Training Dataset

We trained the model on a several sources of dataset provided by ParlAI facebook research (https://github.com/facebookresearch/ParlAI/tree/master/parlai/tasks) and hugging face (https://huggingface.co/datasets/md_gender_bias). However, huge size of original dataset, resulted in a very high processing and training time and memory requirements. Therefore, we had to exclude some sources(like Wikipedia which had 22M rows) and also shrink the remainder dataset to more feasible size for our experimental setting. However, we shrinked each source by the same the proportion of data from each source as in the original. The dataset was processed to match the structure of the evaluation dataset. The final processed dataset had two features: text/sentence and gender label along a particular dimension. Each sample belonged to one of the five class label: ABOUT, AS, TO. Below are list of sources used in our experiment:

Funpedia Funpedia (Milleretal.,2017)[19] contains rephrased Wikipedia sentences in a more conversational way. The dataset only contained biography related sentences. Each conversation was labeled with a gender based on the number of gendered pronouns (he vs. she vs. they). Dinan et al.[2]. There were also texts which had no gender labels or gender was unknown. We excluded these occurrences while training. We used HuggingFace as the source for this data.

Wizard of Wikipedia Wizard of Wikipedia (Dinan et al., 2019c)[18] contains two people discussing a topic in Wikipedia. The dataset only the conversations on Wikipedia biographies and was annotated in the ABOUT dimension Dinan et al.[2].. Again, there were texts for which ABOUT gender was unknown. We excluded these occurrences while training. We used HuggingFace as the source for this data

Yelp The creation of this dataset used Yelp reviewer gender predictor developed by (Subramanian et al., 2018)[17] and retain reviews for which the classifier is very confident about the gender of the content creator of the review(AS dimension) Dinan et al.[2]. We used ParlAI yelp task to load this data. In addition to this, the authors of the paper had also imputed ABOUT label on this dataset using the classifier trained on the ABOUT dimension Dinan et al.[2]. We also used a fraction of this inferred dataset available on HuggingFace as a training sample for ABOUT dimension.

OpenSubtitiles Since the TO dataset for this source was not released by Facebook Research, we used the inferred dataset available on HuggingFace. ABOUT labels were imputed on this dataset by the authors same as the Yelp inferred set Dinan et al.[2]. There were texts which had no gender labels or gender was unknown. We excluded these occurrences while training.

Convai2 ConvAI2 (Dinan et al., 2019b)[20] 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 which we again excluded from the training dataset Dinan et al.[2]. ParlAI Convai2 task was used to load this dataset. We additionally used the inferred dataset on HuggingFace for which ABOUT labels were imputed by the authors same as the Yelp inferred set

LIGHT LIGHT contains persona-based conversation. Similarly to ConvAI2, annotators labeled the gender of each persona (Dinan et al., 2020)[21], giving us labels for the speaker (AS) and speaking partner (TO) Dinan et al.[2]. We additionally used the inferred dataset on HuggingFace for which ABOUT labels were imputed by the authors same as the Yelp inferred set.

Figure 2[2] shows the statistics used by Dinan et al. Figure3 shows the statistics of each dataset we used for our experiments.

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

Figure 2: Original dataset statistics as used by Dinan et al.

Dataset	**M**	**F**	**Dim**
Funpedia	19k	3k	ABOUT
Wizard	6k	1k	ABOUT
Yelp	50k	50k	AS
ConvAI2	22k	22k	AS
ConvAI2	22k	22k	TO
LIGHT	13K	8K	AS
LIGHT	13K	8K	TO
LIGHT Inferred	16K	22K	ABOUT
Yelp Inferred	80K	45K	ABOUT
ConvAI2 Inferred	80K	50K	ABOUT
OpenSub Inferred	70K	30K	ABOUT

Figure 3: Training dataset statistics of the experiment

Dimension	M	F	Total
ABOUT	271K	151K	422K
AS	85K	80K	165K
TO	35K	30K	65K

Figure 4: Total dataset size across each dimension and gender

3.2 Evaluation Dataset

We used the same evaluation dataset, MDGender as in the paper by Dinan Et al., available on HuggingFace(MD_GENDER) and covers all three dimensions. To create this dataset, first involved collecting conversations between two speakers where each speaker is provided with a persona description containing gender information, then tasked with adopting that persona and having a conversation. They are also provided with small sections of a biography from Wikipedia as the conversation topic [2]. In this next phase, a second set of annotators to rewrote 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, what's up? I went for a coffee with my friend and her dog after work as the her indicates a woman. Annotators are additionally asked to label how confident they are that someone else could predict the given gender label. This dataset was collected using crowdworkers from Amazon's Mechanical Turk. Over two thirds of annotators identified as men, which may introduce its own biases into the dataset [2].

Figure2[2] shows the statistics of each dataset we used for our experiments.

4 Experiment

4.1 Software and Hardware

All experiments are run as SLURM jobs on the NYU Greene HPC cluster using NVIDIA RTX8000 and V100 GPUs. We implemented our experiments using PyTorch. We implemented custom train/evaluation data loader using ParlAI facebook research:`https://github.com/facebookresearch/ParlAI/tree/master/parlai/tasks` and HuggingFace:`https://huggingface.co/datasets/md_gender_bias`. For each regularization, we implemented custom loss function for label smoothing, mixout class based on the paper Lee et al. [13], and leveraged NLPaug:`https://github.com/makcedward/nlpaug` for text data augmentation.

4.2 Experimental Setup

For all experiments, we use a batch size of 32 and Adam optimizer. We attempted experimenting with multiple learning rates and report the best metrics in the following section.

4.3 baseline

We used pre-trained distilBERT, a distilled version of BERT(deep bi-directional transformer (Devlin et al).[14]) provided by huggingface transformer library `https://huggingface.co/transformers/model_doc/distilbert.html`). We input the tokenized text data into Distilbert Encoder, and then put the encoded outputs to the multi-layer perceptron, and finally get the log-transformed softmax outputs of 5 classes and compute the loss. Figure3 shows the pictorial description of our model.

4.4 label smoothing

Label smoothing [12] is a technique to avoid overfitting and improve generalization through regularizing a classification model's posterior. It prevents the model from predicting the labels too confidently and mitigate the model uncertainty.

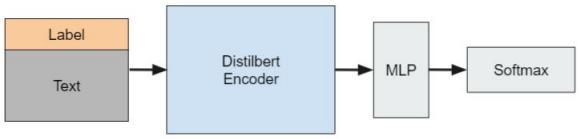

Figure 5: Model diagram

Uniform label smoothing is commonly used in the text classification problems. It modifies the ground truth label by decreasing the probability mass assigned to the ground truth class by some value, τ, and uniformly distributing it across the other labels such that no incorrect label has more probability mass than another, or $\frac{\tau}{n-1}$, where n is the number of classes.

$$y_i^{ULS} = \begin{cases} 1-\tau, & i=j \\ \frac{\tau}{n-1}, & i \neq j \end{cases} \tag{1}$$

While uniform label smoothing uniformly distributes the probability mass that is decreased from the ground truth to the other labels, we propose a variant of label smoothing which spreads the probability mass based on the class similarity between the same gender.

This variant of label smoothing spread the probability mass τ to the classes which has the same gender, where m is the number of classes in the label space with same gender with ground truth jth.

$$y_i^{ULSgender} = \begin{cases} 1-\tau, & i=j \\ \frac{\tau}{[m-1]}, & i \neq j \; and \; i.gender = j.gender \end{cases} \tag{2}$$

We trained our baseline model with label smoothing τ=0.1 and τ=0.2.

4.5 data augmentation

We observe that the training dataset is imbalanced and could lead to imbalanced classification problem. To overcome this, we used data augmentation techniques to balance the training dataset. Wei et al. show that data augmentation of imbalanced datasets can improve accuracy in text classification. We performed two experiments using data augmentation. We first, used text insertion and substitution Wei et al.[16] on the minority classes(TO:Female, TO:Male) to create a less imbalanced dataset. We used the nlpaug library for word replacement and insertion using DistilBERT contextual embedding. We chose 10% of words to be substituted in a sentence. We ran the augmentation for 3 epochs resulting in 6 different samples for each sentence. Generating this data for the entire set of all minority classes was an extremely slow process. So instead we choose to augment the sources which had the lowest number of samples. We trained the model using the training data and the augmented data. In the second instance, we trained the model on the augmented data(insertion and substitution) and also on gender neutral sentences along each dimension. Given that this was multidimensional classifier, the model had to also predict whether the text was about a person/people, talking to a person or talking about self. We aimed to improve the accuracy of classifying a sentence along a dimension(ABOUT, AS,TO) using gender neutral sentences. We predicted that this coupled with balancing the dataset may improve the overall accuracy. We inserted same number of gender neutral sentences for each dimension.

4.6 mixout

Mixout augmentation involves two models, a pretrained model w_{pre} and a current model. Mixout augmentation first applies dropout on the current model, then replace the deleted neurons in the current model with the corresponding parameters in the pretrained model. Since DistilBert contains an embedding part and 5 transform blocks, we apply mixout to all layers in the tranform blocks with a same mixout rate. We experimented with multiple mixout rates(MR), MR = 0.9 and MR = 0.8.

Regularization	Accuracy
Pre-trained DistilBert	16.52%
Cross Entropy (ours)	30.28%
LabelSmooth ($\tau = .2$)	29.68%
LabelSmooth ($\tau = .1$)	30.02%
DataAug (10 % INS,SUB)	31.8%
DataAug (10 % INS, SUB, NEU)	**32.8**%
Mixout (MR=0.9)	30.70%
Mixout (MR=0.8)	30.41%

Table 1: Accuracy of regularization technique variants compared with baseline. INS indicates word insertion, SUB indicated word substitution, NEU indicates addition of Gender Neutral sentences.

5 Result

Table 1 shows the accuracy of each experiment using different regularization technique.

6 Discussion

6.1 Result of label smoothing

The accuracy of the variant of label smoothing was similar to our baseline model or slightly lower. We expected that this label smoothing method would mitigate the model uncertainty(i.e. avoid mispredicting 'asfemale' to 'asmale') and improve the accuracy, but this turned out to be not that effective for our model. From this result, we can generate another hypothesis that this would be because the model is relatively harder to identify dimensionality(AS, TO, About) than gender(male, female). For further analysis, another label smoothing technique which takes both dimensionality and gender into consideration can be devised and then we can identify the effective label smoothing method to mitigate the model uncertainty.

6.2 Result of data augmentation

Through insertion and substitution, we observed a accuracy increase of approximately 1%. We initially augmented only the smallest source of the dataset consisting of the minority labels(To: Female in Convai2) which reported an accuracy of 31. We observed that increasing the number of sources of the dataset(TO:Female, Male in convai2) for augmentation improved the accuracy to 31.8. However,this increase was lower than expected. Though, we augmented the minority classes, the dataset to reduce the imbalance, the dataset was still not perfectly balanced (due to slower augmentation process, we could not augment all minority class sources). Observing the accuracy in both the above mentioned instances, we hypothesise that augmenting all dataset sources with minority classes which will result in a balanced dataset, may improve the overall accuracy considerably. In the second experiment, we observed a better accuracy of 32.8 %. Thus we observed, that improving the accuracy of classifying a sentence along each of the three dimension (ABOUT, AS, TO) may improve the overall accuracy. We again hypothesise that increasing the number of gender neutral sentences will improve the accuracy for dimensionality identification and thus overall accuracy of the gender classifier.

6.3 Result of mixout

The accuracy of different mixout configurations doesn't change much with mixout rate. This might be caused by the relative robustness of the model. We can test with more mixout rates to further analyze the model behaviour.

7 Conclusion

In this project we implemented DistilBert for gender bias problems. We also implemented 3 regularization techniques to further fine tune the model and identified the most effective approach. The

model achieves some progress on the accuracy of the multi-gender bias classification problem, but it is still far from satisfaction. In further work we might train with a larger dataset and Bert model(not DistilBert) and try more regularization methods to achieve a better accuracy. Also, if we could improve the model to achieve high performance from the further experiments, it would be interesting to use this model reversely for detecting and mitigating bias in existing data sets by augmenting data with counterfactuals.

References

[1]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.

[2] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, Adina Williams. 2020a. Multi-Dimensional Gender Bias Classification, *arXiv:2005.00614*.

[3] Alexander H. M., Will F., 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*.

[4]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)*.

[5]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*.

[6]Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Engaging image chat: Modeling personality in grounded dialogue. *arXiv preprint arXiv:1811.00945*.

[7]Pierre Lison and Jorg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles.

[8]Jack Urbanek, Angela Fan, Siddharth Karamcheti,Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktaschel, 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.

[9]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

[10]Audace Nakeshimana and Maryam Najafian (MIT). Case Study with Data: Mitigating Gender Bias on the UCI Adult Database `https://ocw.mit.edu/resources/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/module-four-case-studies/case-study-mitigating-gender-bias/`

[11] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, Simone Teufel. 2020 . It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution *arXiv:1909.00871*

[12]Rafael Muller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help?,2020.

[13]Cheolhyoung Lee, Kyunghyun Cho and Wanmo Kang. Mixout: Effective Reugularization to Finetune Large-scale Pretrained Language Models. *arXiv:1909.11299*.

[14]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language under- standing. *arXiv:1810.04805*

[15] Victor Sanh, Lysandre Debut, Julien Chaumond and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs.CL]*

[16] Jason Wei, Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks

[17] Sandeep Subramanian, Eric Michael Smith, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. arXiv preprint arXiv:1811.00552.

[18] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019c. Wiz- ard of Wikipedia: Knowledge-powered conversa- tional agents. In Proceedings of the International Conference on Learning Representations (ICLR).

[19] Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research soft- ware platform. In Proceedings of the 2017 Con- ference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 79–84, Copenhagen, Denmark. Association for Computa- tional Linguistics.

[20] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019b. The second conversational intelligence challenge (ConvAI2). arXiv preprint arXiv:1902.00098.

[21] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020b. Queens are powerful too: Mitigating gender bias in dialogue generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)