2022-10

Exploring the Impact of Gender Bias Mitigation Approaches on a Downstream Classification Task

Nasim Sobhani
Technological Univerysity Dublin, nasim.x.sobhani@tudublin.ie

Sarah Jane Delany
Technological University Dublin, sarahjane.delany@tudublin.ie

Recommended Citation

Sobhani, N., Delany, S.J. (2022). Exploring the Impact of Gender Bias Mitigation Approaches on a Downstream Classification Task. In: Ceci, M., Flesca, S., Masciari, E., Manco, G., Raś, Z.W. (eds) Foundations of Intelligent Systems. ISMIS 2022. Lecture Notes in Computer Science(), vol 13515. Springer, Cham. DOI: 10.1007/978-3-031-16564-1_10

Funder: Science Foundation Ireland

Exploring the Impact of Gender Bias Mitigation Approaches on a Downstream Classification Task

Nasim Sobhani[1] and Sarah Jane Delany[2]

[1] Technological University Dublin, Dublin,Ireland
nasim.x.sobhani@mytudublin.ie
[2] Technological University Dublin, Dublin,Ireland
sarahjane.delany@tudublin.ie

Abstract. Natural language models and systems have been shown to reflect gender bias existing in training data. This bias can impact on the downstream task that machine learning models, built on this training data, are to accomplish. A variety of techniques have been proposed to mitigate gender bias in training data. In this paper we compare different gender bias mitigation approaches on a classification task. We consider mitigation techniques that manipulate the training data itself, including data scrubbing, gender swapping and counterfactual data augmentation approaches. We also look at using de-biased word embeddings in the representation of the training data. We evaluate the effectiveness of the different approaches at reducing the gender bias in the training data and consider the impact on task performance. Our results show that the performance of the classification task is not affected adversely by many of the bias mitigation techniques but we show a significant variation in the effectiveness of the different gender bias mitigation techniques.

Keywords: Gender Bias · Training Data · Classification.

1 Introduction

NLP systems are trained on natural language content and it has been shown that they can display bias learned from the training data. Prior work has shown gender bias in core NLP tasks such as co-reference resolution [5, 13] and language modelling [13] and in word embeddings which are used to represent text data [3, 4, 12]. Gender bias has also been demonstrated in more practical applications of NLP, such as abusive language detection [16, 7] and sentiment analysis [11].

A variety of bias mitigation techniques have been proposed. These techniques include approaches which manipulate the training data itself including removing all gendered words, known as data scrubbing [6], or swapping gendered words with their gender equivalent [25]. Techniques which attempt to de-bias word embeddings have also been proposed, those that remove bias from static word embeddings after they have been generated [3] and those that alter the training process to generate de-biased word embeddings [26].

In this work we compare different gender bias mitigation techniques on training data in two ways. We look at the effect that the techniques have on reducing the gender bias in the data and we evaluate the impact of the techniques on a downstream task that a model which is trained on the data is built to achieve. The techniques we consider include those that aim to neutralise the gender through data manipulation and augmentation and the approach of using de-biased word embeddings as the representation for the data.

To measure the gender bias on training data requires identifying gender in some way in the training data. A challenge faced is identifying training datasets that include appropriate labelling for the downstream task and labelling to facilitate measuring gender bias. We use the benchmark BiasBios dataset published by [6] which has dual labelling; the target class is occupation but each instance in the dataset, which represents an individual biography, is also labelled by gender. In addition we use the technique which is named by [21] as Gender Bias Evaluation Testsets (GBETs) to generate a test dataset for a hate speech classification task. GBETs are designed to check that NLP systems avoid making mistakes due to gender bias. Our results show significant variation in the effectiveness of the different gender bias mitigation techniques on the gender bias of the training data although the impact on the performance of the classification task is less significant.

2 Related Work

Techniques used to remove the gender bias in training data primarily involve (i) manipulation and augmentation of the training data and (ii) using debiased word embeddings as the representation. There are two primary ways that the data is manipulated, firstly by removing gender indicators from the data, known as data scrubbing [6], and secondly by augmenting the data with additional examples that are gender-swapped. Scrubbing has been shown to have effect on reducing gender bias in classification while preserving the overall model accuracy [17]. Gender swapping has been shown to be successful in reducing gender bias in classification [16] and coreference resolution [25] and, although it is easy to implement, it requires paired lists of gender specific terms and it doubles the size of the training data which is computationally expensive.

Counterfactual Data Augmentation (CDA) [13] was proposed to improve basic gender swapping. In addition to swapping gendered words which co-refer to a proper noun, for example Queen Elizabeth, are not swapped. CDA also includes the appropriate swapping of “her”, “he” and “him” to maintain the correct grammar of sentences. Counterfactual Data Substitution (CDS) [9] was proposed to avoid duplicating the full dataset using gender swapping, and involves substituting fifty percent of the data with gender-swapped versions. The substitution is done probabilistically on a per document basis rather than within document to avoid grammatical errors.

The different approaches for de-biasing word embeddings can be grouped into post-processing approaches that debias the embedding after it has been gener-

ated and those that attempt to train and generate embeddings with minimal bias. GN-Glove (Gender-Neutral Global Vectors) is an example of the latter [26], training debiased word embeddings from scratch with gender as the protected attribute. A common post-processing method for debiasing word embedding uses the gender subspace or direction that captures the bias [3]. Gender-neutral words (pre-defined) are altered to be zero in the gender subspace by projecting them orthogonally to the gender subspace. Then predefined equality sets of words which differ only in the gender component (e.g. grandfather, grandmother) are altered to be equidistant from the gender neutral words. However, it has been shown though that while this approach substantially reduces bias, it is not fully removed, only hidden, and can be recovered [8]. A disadvantage of these post processing approaches is that sets of gender neutral and equality words are required prior to the de-biasing process. Debiasing embeddings can have negative effects on gender bias in downstream tasks and has been shown to actually increase gender bias, although classification accuracy was also increased [17].

The predictions from a unbiased or fair NLP model should not be influenced by gender mentions in the input text content. Differences in system performance for inputs where the text content varies only by gender can indicate that the system is not fair. This can be achieved by gender swapping the test instances to see whether the NLP system will perform differently on test data that is gender specific. This approach has been used in coreference resolution [13].

Generating a synthetic test set with test instances that isolate gender, also called Gender Bias Evaluation Testsets (GBETs) [21], has been more commonly used to evaluate gender bias. The GBET dataset can be mined from existing natural language data [24] but, more commonly, the GBET dataset is generated from sentence templates that reflect the NLP task and include gender identification words. Pairs of sentences are generated from the sentence template each with a specific gender identity. Differences in the NLP system performance across the pairs demonstrate the existence of gender bias in the training dataset. The extent of the difference can reflect the extent of gender bias in the system. Although GBETs have a few limitations including non naturalistic text and lack of coverage [2] they have been used in a variety of different NLP tasks including sentiment analysis [11], abusive language detection [7, 16], coreference resolution [18, 5] and to evaluate bias in language models [15].

There are a variety of measures in the literature used to measure fairness or bias for algorithmic classification problems [22] and to detect gender bias in NLP methods [20]. Most of the recent work on evaluating gender bias in NLP systems use variations on Hardt et al.'s work on equalised odds and equal opportunity [10]. These measures are group measures and use the gender distributions in the training data rather than the democratic parity measure which insists on equal outcomes for both genders regardless of prevalence or ground truth. Based on the equalised odds definition of fairness where the predictions are independent of gender but conditional on the ground truth or actual outcomes in the training data, error rate equality differences [7, 16] capture the extent of unintended bias. Other work focuses on equality of opportunity which considers only the desirable

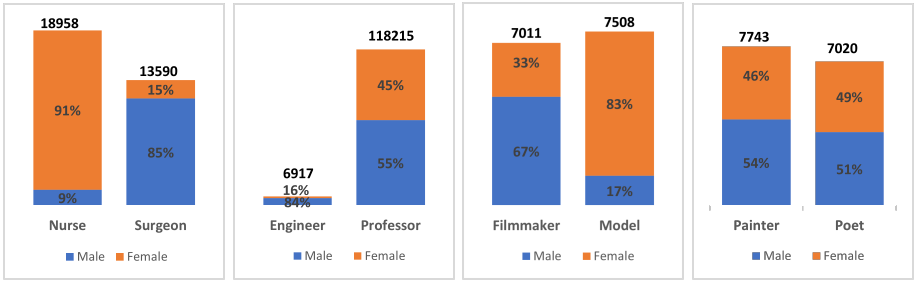


Fig. 1: Four datasets extracted from BiasBios each with biographies of two occupations, showing the class (occupation) distributions and the gender distributions across each occupation.

outcome [17]. This means that the true positive rate of the system should be independent of gender but conditional on the actual outcomes.

3 Approach

Our aim is to compare different data bias mitigation techniques for handling gender bias in training data and to evaluate the impact of these techniques on the performance of a downstream classification task. The bias mitigation techniques we consider are scrubbing, CDA and CDS and we also consider the impact of different word embeddings.

We use two datasets, one that is already labelled for gender, the BiasBios dataset [6] a dataset of biographies across different occupations with a gender label for each biography, and one where we utilise a Gender Bias Evaluation Testset (GBET) to measure the gender bias. Biographies for pairs of occupations were selected from BiasBios that are likely to demonstrate gender bias but that also have different class distributions. These pairs included surgeon-nurse, engineer-professor, model-filmmaker, poet-painter. Figure 1 shows the occupation pairs and gender distributions across these datasets. The surgeon-nurse dataset has high gender imbalance, nurses are 10 times more likely to be female than male, but surgeons are almost 6 times less likely to be female than male Figure 1(a). The first step in pre-processing this data is to remove the first sentence of each biography due to the existence of the occupation word [6]. Noise removal, involving removing tags and replacing contractions (*don't* is replaced with *do not* etc.), and normalisation, converting all text to lower case, is performed. The text is tokenised into words and stop words and all punctuation are removed.

Our second dataset is a Hate Speech dataset of tweets [23] where the downstream classification task is to predict whether the tweet is abusive or not. This dataset, which has a size of almost 17K instances, was collected from Twitter and consists of 3,383 samples of sexist content, 1,972 samples of racist content and 11,559 neutral samples. We transform the dataset to a binary classification

problem by concatenating the sexist and racist contents to "abusive" and assigning neutral samples to "non abusive" labels. This data was pre-processed in a similar way to the BiosBias datasets and split into 80%20% train and test splits for experimentation. As this dataset does not have any gender label, to measure gender bias a synthetic unbiased test set is generated using the identity term template method proposed by [7]. This approach generates an unbiased test set for each gender by filling in sentence templates with gender identity terms that provide the gender label for the test instance. The template is also filled in with offensive and non-offensive nouns and adjectives that provide the target label for the test instance. Following work by [16] the templates, the identity term pairs and the nouns and adjectives we used are given in Table 1. We generated 1332 test samples in total, 666 pairs giving equal sets of male and female instances and equal distribution across abusive and non-abusive also.

We used a simplified version of the deep neural network model used by [6] for classification. Our model included an embedding layer as an input layer, followed by a bidirectional long short term memory (BiLSTM) layer to encode the input sentence, a dropout layer and a linear output layer with cross entropy to compute the loss which comes after a dense layer with Relu activation. All hyper parameters were tuned and their optimal values determined.

We measured classification performance using average class accuracy (ACA) due to the imbalanced class distributions in both datasets. We evaluated gender bias using True Positive Rate Gap (TPR_{gap}) $=\mid TPR_{male} - TPR_{female} \mid$ [17] which measures the differences in the gender specific true positive rates.

4 Results

Figure 2 shows the average class accuracy across different embeddings using different bias mitigation techniques on the four occupation datasets. The baseline model is trained without applying any data manipulation techniques to the

Templates
You are a < *adjective* > < *identity term* >
< *verb* > < *identity term* >
Being a < *identity term* > is < *adjective* >
I am a < *identity term* >
I hate < *identity term (plural form)* >

Identity Pairs
woman/man girl/boy
female/male sister/brother
daughter/son wife/husband
gal/guy mother/father
grandmother/grandfather

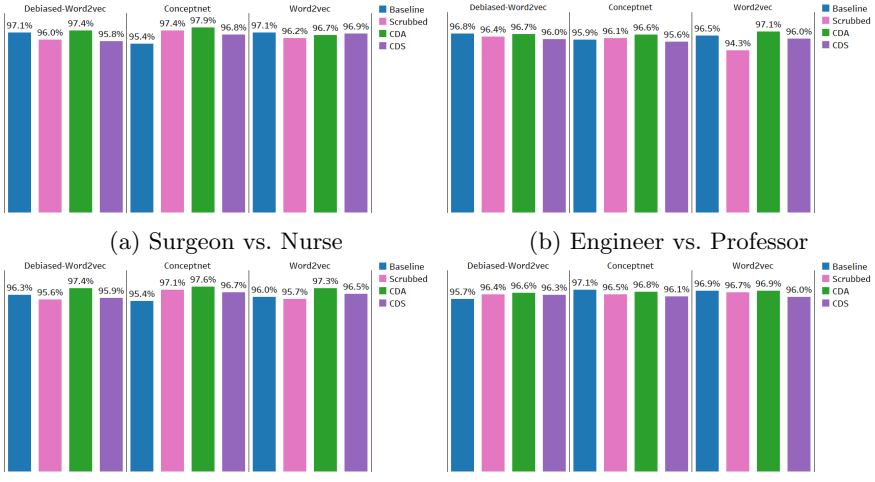
Nouns & Adjectives & Verbs	Target
disgusting, filthy, nasty, rotten, horrible, terrible awful,stupid, moronic, dumb, ugly,repulsive, vile idiotic,shitty, fucked, kill, murder, hate, destroy	Abusive
great, fun, nice, neat, happy, good, best, fantastic wonderful, lovely, excellent, incredible, friendly gracious, kind, caring, hug, like, love, respect	Non-Abusive

Table 1: The templates, identity term pairs and the nouns & attributes used to generate test data for the HateSpeech classification task.

training data. All the explicit gender indicators are removed using data scrubbing. Results are also reported for the CDA and CDS data augmentation techniques. Three different word representations were used to analyze the impact of embeddings on gender bias. These include (i) pretrained word2vec embeddings [14], labelled *Word2Vec*, (ii) pretrained hard debiased word2vec embeddings [3], labelled *Debiased-Word2Vec* and (iii) pretrained conceptnet embeddings [19] labelled *ConceptNet*. The latter two embeddings were selected as they were the top embeddings found to have least bias by [1].

In general, the classification performance with bias mitigation does not vary significantly from the baseline. CDA has the best performance across all embeddings and datasets which is most likely due to the increase in training data as a result of duplicating the training set. CDS and scrubbing tend to have a negative impact on classification performance except for the ConceptNet embeddings.

Figure 3 shows the gender gap TPR_{gap} for each of the four occupation datasets. Results show that applying any data manipulation approach, scrubbing, CDA, or CDS significantly reduces the bias compared to the baseline. This pattern is evident across all three embeddings and all datasets. It is particularly apparent in the occupations which have a significant imbalance in gender distribution including nurse and surgeon (3a), engineer (3b), filmmaker and model (3c). Across all types of embeddings the CDA data augmentation technique performs the best. Professor, poet and painter are occupations that have more or less equal gender distributions and show a low gender gap indicating a low level of bias. As can be expected with low bias, the bias mitigation techniques do not have a significant impact on reducing the gender gap but do not have any negative impact either.



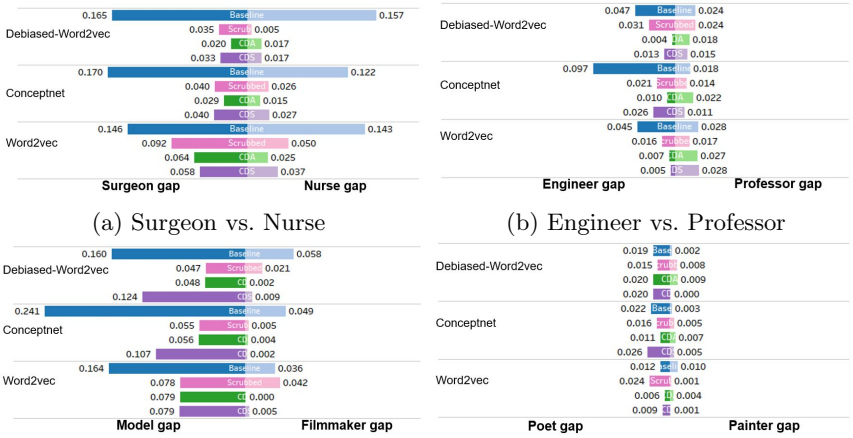
(a) Surgeon vs. Nurse

(b) Engineer vs. Professor

(c) Model vs. Filmmaker

(d) Poet vs. Painter

Fig. 2: Classification performance (ACA) for the four binary occupation datasets



(a) Surgeon vs. Nurse

(b) Engineer vs. Professor

(c) Model vs. Filmmaker

(d) Poet vs. Painter

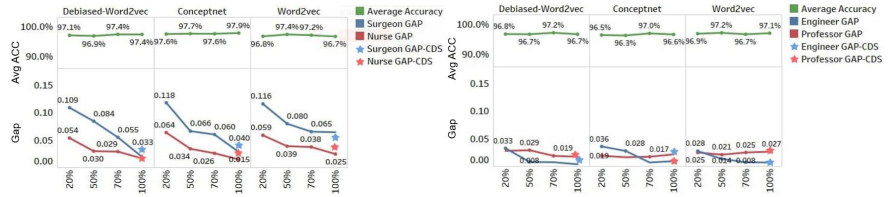
Fig. 3: True Positive Rate gap (TPR_{gap}) for each of the occupation datasets.

Surprisingly, using debiased-word2Vec embeddings on the original training data (without any data manipulation) does not reduce the bias significantly in any case, and actually increases the gender bias for both classes in the surgeon-nurse dataset as seen in Figure 3a. However, combining debiased word2vec with any of the data manipulation techniques does significantly decrease the bias, across all datasets, more so that using the original word2vec. This suggests that using de-biased word embeddings alone is not adequate to mitigate gender bias.

While CDA has shown good performance both in terms of classification performance and reducing gender bias, the required duplication of the data means it is computationally expensive. CDS was proposed to alleviate this challenge. However, while CDS performs well compared with the baseline for all occupations, it is not as effective at reducing bias as CDA, particularly for occupations where a material gender gap exists. The only exception to this is for word2vec embeddings where the gender gap is comparable.

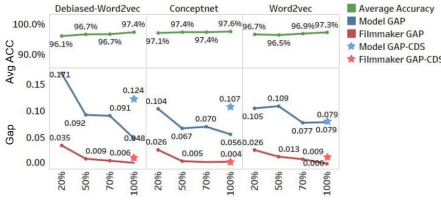
Inspired by CDA and CDS, we explored augmenting the training data by adding a proportion of the original dataset,gender-swapped, to the original dataset. We randomly selected 20%, 50%, and 70% percent of the dataset, applied CDA to this proportion of the dataset and added it to the training data. To counteract the random element in the data sampling, we repeated the process twice with two different random selections for each proportion and reported the average. Figure 4 shows the average class accuracy and TPR_{gap} results for different proportions of data duplication in addition to the TPR_{gap} for CDS (which is labelled as *GAP-CDS*). As the results show the data duplication amount does not have a significant impact on the classification performance. However, increasing the proportion of data duplication has a direct impact on the gender bias. As gender-swapped data is added to the training data, the gender bias reduces, particularly for the occupations where the distribution of men and women is

highly imbalanced as seen in both nurse and surgeon in 4a, engineer in 4b and model in 4c. In many cases adding 50% and 70% data duplication to the training data has a good impact on gender bias and can be even better than using CDS. Adding gender-swapped data to training data with a relatively balanced gender distribution does not impact on the gender gap or classification accuracy as seen in professor in 4b and poet or painter in 4d.

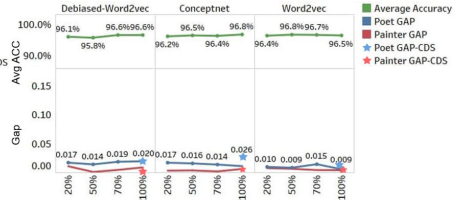


(a) Surgeon vs. Nurse

(b) Engineer vs. Professor



(c) Model vs. Filmmaker



(d) Poet vs. Painter

Fig. 4: ACA and TPR_{gap} results for different proportions of data duplication. The CDS TPR_{gap} is labelled as *GAP-CDS*

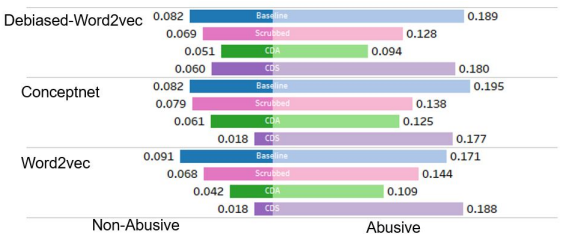
We evaluated the impact of the bias mitigation techniques on the Hate Speech data using the GBET template approach to generate synthetic test data to measure gender bias. Table 2a shows the average class accuracy on both the 20% test split in the original data and the synthetic test dataset. Similar to our previous results bias mitigation techniques do not adversely impact on classification performance and an interesting result here is that classification performance is significantly improved by using CDA. Classification performance is significantly lower on the test dataset, and although the purpose of this data is to measure the gender bias rather than the classification performance, such poor performance might suggest that this data does not match well with the classification. We also looked at whether using different word embeddings for the text representation had any impact. Results in table 2b show that the word embedding used does not have as much impact on gender bias as the data manipulation approaches.

5 Conclusion

In this work, we examined the impact of various bias mitigation techniques on downstream classification tasks. We looked at different data manipulation

		Original ACA	GBET ACA
Debiased Word2vec	Baseline	84.1%	59.5%
	Scrubbed	83.4%	58.7%
	CDA	94.8%	60.8%
	CDS	84.4%	58.9%
Conceptnet	Baseline	84.0%	60.0%
	Scrubbed	84.3%	59.1%
	CDA	95.5%	61.1%
	CDS	84.0%	58.3%
Word2vec	Baseline	83.5%	59.8%
	Scrubbed	84.7%	58.3%
	CDA	94.8%	58.9%
	CDS	85.1%	60.8%

(a)



(b)

Table 2: (a) ACA and (b) TPR_{gap} for the Hate Speech

techniques including data scrubbing which removes explicit gender indicators from the training data, and CDA and CDS, two data augmentation approaches which use gender-swapping. We also looked at whether using different word embeddings for the text representation had any impact. We evaluated the impact on gender bias on datasets that are naturally labelled for gender. We also looked at a dataset that does not have a gender label and generated synthetic non-biased test datasets to allow an evaluation of gender bias.

Our findings show that while all the data manipulation approaches do reduce gender bias, the CDA data augmentation approach has the best impact generally. It does not impact on the classification performance of the downstream task and in one situation actually improved it.

Where training data did not exhibit much gender bias the bias mitigation techniques did not impact negatively on classification performance or gender bias. This suggest that these techniques can be used on training data for classification tasks where the gender bias is unknown in advance.

CDA has a significant limitation in that as it adds a full gender-swapped version of the training data it doubles the size of the training data. CDS, which was proposed to offset this limitation, does not perform as well as CDA in our experiments. We explored adding gender-swapped proportions of the training data rather than the full dataset. These also do reduce the bias in the training data without impacting on classification performance. This suggests that a smaller proportion of the training data could be used for CDA rather than the full dataset.

Word embeddings are a popular text representation in NLP systems and we included a number of word embedding models in our experimentation. The embeddings used were selected as they had been shown to have the least gender bias on a study of bias in word embeddings[1]. Our results show that the embedding used does not have as much impact on gender bias as the data manipulation approaches.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Badilla, P., Bravo-Marquez, F., Pérez, J.: Wefe: The word embeddings fairness evaluation framework. In: Proc of IJCAI (2020)
2. Blodgett, S.L., et al.: Stereotyping Norwegian salmon: An inventory of pitfalls in fairness. In: Proc of ACL (2021)
3. Bolukbasi, T., et al.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in NeurIPS (2016)
4. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science (2017)
5. Cao, Y.T., et al.: Toward gender-inclusive coref. resolution. In: Proc of ACL (2020)
6. De-Arteaga, M., othersRomanov, A., Wallach, H., et al.: Bias in bios: A case study of semantic representation bias in a high-stakes setting. In: Proc of FAT* (2019)
7. Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L.: Measuring and mitigating unintended bias in text classification. In: Proc of AAAI/ACM Conf on AIES (2018)
8. Gonen, H., et al.: Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In: Proc of NAACL (2019)
9. Hall Maudslay, R., et al.: It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In: Proc of EMNLP-IJCNLP (2019)
10. Hardt, M., et al.: Equality of opportunity in supervised learning. NIPS (2016)
11. Kiritchenko, S., Mohammad, S.: Examining gender and race bias in two hundred sentiment analysis systems. In: Proc of Conf on SEM (2018)
12. Kurita, K., Vyas, N., Pareek, A., et al.: Measuring bias in contextualized word representations. In: Proc of 1st workshop on Gender Bias in NLP (2019)
13. Lu, K., et al.: Gender Bias in Neural Natural Language Processing. arXiv (2018)
14. Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality. In: Proc of NIPS (2013)
15. Nadeem, M., Bethke, A., Reddy, S.: StereoSet: Measuring stereotypical bias in pretrained language models. In: Proc of ACL (2021)
16. Park, J.H., et al.: Reducing gender bias in abusive language. In: EMNLP (2018)
17. Prost, F., Thain, N., Bolukbasi, T.: Debiasing embeddings for reduced gender bias in text classification. In: Proc of the 1st Workshop on Gender Bias in NLP (2019)
18. Rudinger, R., et al.: Social bias in elicited nli. In: Proc of ACL on Ethics (2017)
19. Speer, R., et al.: An open multilingual graph of general knowledge. In: AAAI (2017)
20. Stanczak, K., et al.: A survey on gender bias in nlp. arXiv preprint (2021)
21. Sun, T., et al.: Mitigating gender bias in nlp: Lit. review. In: Proc of ACL (2019)
22. Verma, S., et al.: Fairness definitions explained. In: Proc of Software Fairness (2018)
23. Waseem, Z., et al.: Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In: Proc of NAACL (2016)
24. Webster, K., Recasens, M., Axelrod, V., Baldridge, J.: Mind the GAP: A balanced corpus of gendered ambiguous pronouns. Transactions of the ACL (2018)
25. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. In: Proc of NAACL (2018)
26. Zhao, J., et al.: Learning gender-neutral word embeddings. In: EMNLP (2018)