Mitigating Gender Bias in Text Classification using a Graph Convolutional Network

Butt, Natasha
Dept of Statistical Science, UCL
natasha.butt.19@ucl.ac.uk

Dellaporta, Charita
Dept of Computer Science, UCL
charita.dellaporta.19@ucl.ac.uk

Dobrowolska, Agnieszka
Dept of Computer Science, UCL
aga.dobrowolska.16@ucl.ac.uk

Meier, Johanna
Dept of Statistical Science, UCL
johanna.meier.19@ucl.ac.uk

Abstract

This paper extends the work of De-Arteaga et al. (2019) to quantify and mitigate gender bias in occupation classification of biographies, performed using a Graph Convolutional Network (GCN). Our GCN predictions are found to have less bias than the predictions generated by De-Arteaga et al. (2019) with bag-of-words, fasttext word embeddings, and deep recurrent neural networks. We aim to debias our training corpus by removing explicit gender indicators from the biographies and find that the bias in the predictions from our re-trained GCN is reduced. By excluding occupations from our classification task which are not well represented in our dataset, we also find a bias reduction in the predictions.

1 Introduction

The recent years have seen a large increase in the adoption of Natural Language Processing (NLP)-based machine learning methods for a growing number of tasks such as automated decision making, recommendation tasks and reading comprehension. As the presence of these systems becomes increasingly ubiquitous in our everyday lives, they move from being merely passive systems to having a more active effect on society, influencing which media articles appear on a feed or which job adverts one sees. However, there is increasing evidence that some outputs of these models may propagate biases which already exist in today's society. If allowed to go unchecked, decision-making based on these machine learning systems may contribute to discrimination.

One area in which these concerns are of particular importance is that of recruitment. As the hiring process is increasingly moving online, new ways are constantly sought to automate processes such as deploying NLP-based methods for target-advertising of job postings. This requires the model to determine candidate's current job position, their skills, interests and their 'potential' for the new role. These factors are then taken into account to determine whether the ad will be shown to the candidate. However, the seemingly simplest of these tasks - determining the candidate's current occupation - proves to be challenging. De-Arteaga et al. (2019) find evidence that the occupation classification process can propagate gender bias arising from the real-world occupation gender imbalances that are reflected in the training data.

This paper aims to extend the work of De-Arteaga et al. (2019) to quantify and mitigate the gender bias when occupation classification is performed using a GCN. We predict a candidate's occupation from a given biography and experiments are performed to quantify the gender bias and assess whether removing gender indicators from the biographies, such as names and pronouns, is enough to mitigate this gender bias.

2 Related Work

2.1 Mitigating Gender bias in NLP

Related work investigates reducing gender bias in NLP tasks through adjusting algorithms as well as debiasing word embeddings and training corpora.

Adjusting algorithms: Zhao et al. (2017) impose constraints on optimisation during training,

derived from the gender distribution in the training corpus. Gender bias amplification is defined as when model predictions have more bias than is present in the training data. However, we note that even if amplification is removed, predictions may still have bias due to bias in the training corpus. In comparison, Zhang et al. (2018) focus on mitigating bias, not bias amplification, through adversarial learning and employ a wider range of bias measurements. Here, the adversary aims to model a protected variable such as gender. However, the adversarial training method is very sensitive to the chosen hyperparameters, often leading to divergence. Therefore, overall progress is underway in developing adjusting algorithms that mitigate bias and bias amplification. However, this is a relevantly nascent area and significant challenges remain for wider applications throughout NLP tasks.

Debiasing word embeddings: Another related field of research looks at debiasing word embeddings in single class Bolukbasi et al. (2016) and multiclass Manzini et al. (2019) settings. The aim is to remove gender related bias such as the association between words like *surgeon* and *male*, while preserving useful information such as the association between *mother* and *female*. However, this work relies on words that represent bias (referred to as defining sets) and words that should or should not contain bias (referred to as equality sets). Therefore, there is subjectivity and thus the sets may not fully capture the bias subspace. Further, Gonen and Goldberg (2019) argue these methods only manage to 'cover up' the bias as in most cases bias can still be recovered from the distances between words in the embedding space. In summary, there is significant study in debiasing word embeddings, however, there are open questions surrounding the appropriateness for mitigating bias in more generalised NLP settings.

Debiasing training corpora: A substantial field of study also investigates debiasing training corpora. Our work closely follows that of De-Arteaga et al. (2019) who aim to debias training corpora by removing explicit gender indicators. This successfully reduces bias without affecting performance. However, only three semantic representations are used: bag-of-words, word embeddings and deep recurrent neural networks (RNNs). Further, it would be interesting to see how these classifiers look with smaller datasets, reflecting those that small recruitment firms may have access to. Other examples of successfully debiasing training corpora include Zhao et al. (2018) in coreference resolution and language modelling tasks. The dataset used is a unification of their original dataset and one in which the gender indicators have been swapped, e.g. *she* to *he*. However, not only may this come at significant computational and memory cost, this new dataset may no longer accurately reflect real life and so may reduce the reliability of results. Furthermore, some sentences may lose their meaning; for example, "she gave birth" in the original dataset would result in "he gave birth" in the new dataset.

Our work extends research into debiasing training corpora by re-training our model with explicit gender indicators removed and comparing predictions. Swapping is considered for analysis purposes. Our novel use of a GCN in this area enables comparisons of gender bias for different classifiers in De-Arteaga et al. (2019), contributing more widely to the detection of gender bias in NLP.

2.2 Word Representation for Text Classification

A crucial consideration for text classification is the choice of representation. In their work, De-Arteaga et al. (2019) explore a bag-of-word representation, `fasttext` word embeddings and a representation generated using a deep RNN. When using pre-trained word embeddings, the performance of the model largely depends on how 'good' the embeddings are. There exist various approaches for obtaining document word embeddings, such as obtaining individual word embeddings and aggregating them to produce document representations. However, the main drawback here is that the document representations are built after learning the word representations. The alternative approach of learning text representations using convolutional neural nets (CNNs) and RNNs is shown to be quite effective and is widely used, however, they focus on learning localised sequences of consecutive words. Therefore, they do not explicitly utilise information such as global word co-occurrence and hence may not capture the global structure of the dataset very well.

Here, we extend the work of De-Arteaga et al. (2019) by considering representations generated using a Graph Neural Network (GNN), as they are able to capture information about the global graph structure in the graph embeddings. GNNs can

be thought of as generalisations of standard neural network models, such as CNNs. While CNNs are designed to work on regular grid structures, such as a 2-dimensional mesh or a 1-dimensional sequence, GNNs are able to take an arbitrarily-structured graph as input. Hence, a graph convolution is a generalisation of 2D convolution, which is essentially the weighted average of a node's neighbourhood. For instance, an image can be thought of as a special case of a graph where each node (pixel) is connected to its adjacent nodes. However, in contrast to an image, the nodes in a graph are un-ordered and unequal in size. As the sheer computational expense of building and storing a large graph in RAM is very significant, in this work we specifically focus on a GCN, a simplified GNN model recently introduced by Kipf and Welling (2016), which nonetheless was shown to achieve state-of-the-art results.

There exist different ways of generating text representations using GCNs. In the recent years, GCNs have been applied on parse trees to generate word representations capturing syntactic dependencies. This is desirable as syntactic representations are closely related to semantic representations, however previous attempts without utilising GCNs suffered from massive vocabulary-size problems. Here, node dependencies on the graph can represent the dependency context of words without having to increase the vocabulary size.

Marcheggiani and Titov (2017), Li et al. (2018) and Bastings et al. (2017) used GCNs syntactic encoders in this way for semantic role labeling, clinical notes classification and machine translation tasks respectively. All three of the papers found that GCNs were better at capturing long distance dependencies than other neural models, such as LSTMs. Though Vashishth et al. (2019), who used GCNs for document dating, treating it as a classification problem and adapting the work by Kipf and Welling (2016) for directed graphs, observed that best results are obtained by combining GCNs with models that capture local sequential dependencies, such as LSTMs. Li et al. (2018) also noted that while GCNs can capture context variations in a large corpora more effectively than sequence-aware models, they also require larger training datasets for parameter estimation.

Recently, Yao et al. (2018) proposed a novel text classification approach using GCNs by converting the entire corpus into a single text graph, constructed using word co-occurrence and document relations. The method proposed by Yao et al. (2018) explicitly utilises information about the global structure of the corpus and learns the word and document embeddings jointly, thus capturing more latent information. None of the previously explored methods are capable of this. The documents and words are first represented using one-hot vectors and then embeddings are learned jointly by the GCN in a supervised manner using the documents' class labels. Thus, the task of text classification is converted into a task of node classification. Our work builds on this approach by measuring and mitigating gender bias introduced when classifying text using this method.

3 Methods

Here, we will detail the GCN methodology as derived by Kipf and Welling (2016) and adapted by Yao et al. (2018), which comprises a spectral-based GNN, as opposed to a spatial one (Wu et al., 2019). Formally, we define a graph as $G = (V, E)$ where V is a set of vertices or nodes and E is a set of edges, such that $\mid V \mid = n$. We let $v_i \in V$ denote a node and $e_{ij} = (v_i, v_j) \in E$ an edge between v_i and v_j. Every node is assumed to be connected to itself, thus $(v_i, v_i) \in E \ \ \forall i$. All the nodes with their features are stored in a matrix $X \in \mathbb{R}^{n \times m}$, where m is the dimension of the feature vectors. The adjacency matrix A of G is a $n \times n$ matrix such that $A_{ij} = 1$ when $e_{ij} \in E$ and $A_{ij} = 0$ when $e_{ij} \notin E$. This definition, alongside with assuming that each node is connected to itself, ensures that the diagonal elements of A are set to 1. This is essential so that the aggregated representation of the node's neighbourhood includes the representation of the node itself. We also define the degree matrix of G, D, where $D_{ii} = \sum_j A_{ij}$. A single layer convolution in a GCN can aggregate only information about the immediate neighbours of a given node. Thus, stacking multiple GCN layers is essential and results in aggregating information from nodes that are an increasing number of hops away. A generalised computation of the GCN outputs a new node feature matrix L^{j+1} taking the form of:

$$L^{j+1} = \rho\left(\tilde{A} L^j W_j\right)$$

where $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ denotes the normalised symmetric adjacency matrix, W_j is the weight matrix corresponding to this layer and L^j, is the output matrix from the previous layer, with $L^0 = X$.

ρ corresponds to the activation function - in this project we used ReLU: $\rho(x) = \max(0, x)$. The objective function used here is the cross-entropy loss over all labelled biographies, such that

$$\mathcal{L} = -\sum_{d \in \mathcal{Y}_D} \sum_{f=1}^{F} Y_{df} \ln Z_{df}$$

where Y_{df} is the set of biography indices, F is the dimension of the output features corresponding to the number of classes, Y_{df} is the binary label indicator matrix and Z is the probability of the document d belonging to class f, which is the output of the model passed through a *softmax* classifier.

3.1 Graph Convolutional Networks for Text Classification

The graph, which the GCN will work on, is built from a subset of the BiosBias dataset. This dataset is converted into a single graph, where the nodes correspond to words and biographies. Hence, the number of nodes $\mid V \mid$ is equal to the number biographies plus the number of words in the vocabulary. Since we want to learn the word and document representations, the words should be represented using one hot vectors. Thus, we find that $X = I$, i.e. the feature matrix is simply the identity matrix. The **word-document** edges edges are first built in a binary way using occurrence information: an edge exists if the word occurs in the given document. These edges are then weighted with TF-IDF (term frequency - inverse document frequency), where TF is the number of times a word occurs in a given document, while IDF is the inverse fraction of the number of documents that contain that word, scaled logarithmically. The benefit of using TF-IDF over just TF is that it captures some information about the 'rareness' of the term and it was empirically found to outperform using just TF. To obtain global co-occurrence information a sliding window was applied to all the documents in the corpus. Meanwhile, the **word-word** edges are computed using PMI (Point-wise Mutual Information) which compares the joint and marginal probability distributions of the given words, and thus measures the probability of their coincidence, assuming independence. Therefore, a high positive value of PMI suggests that the words are semantically similar, while a negative value implies little or no semantic similarity. Accordingly, word-word edges are only built for word pairs that have positive PMI. The above formulation is summarised below:

$$A_{ij} = \begin{cases} \text{TF} - \text{IDF}_{ij} & i \text{ is document , } j \text{ is word} \\ \text{PMI}(i,j) & i, j \text{ are words,} \\ & \text{PMI}(i,j) > 0 \\ 1 & i = j \\ 0 & \text{otherwise} \end{cases}$$

The PMI value of two words i, j is computed as:

$$\text{PMI}(i,j) = \log \frac{p(i,j)}{p(i)p(j)}$$

where $p(i,j) = \frac{|W(i,j)|}{|W|}$ and $p(i) = \frac{|W(i)|}{|W|}$, such that $\mid W \mid$ is the total number of sliding windows, $\mid W(i) \mid$ is the number of sliding windows that contain word i and $\mid W(i,j) \mid$ is the number of sliding windows that have both words i and j.

It is worth noting that here document-document edges are not required since we built the entire corpus using a single heterogeneous graph. Thus it is not necessary to approximate relations between individual documents which in turn should improve the performance of the model.

3.2 Implementation Details and Baselines

For all of the models, first the dataset was cleaned and tokenized, following the work of Kim (2014). Next, the stop words were removed - here, it was essential that initially all the gender indicators were retained, thus they were manually removed from the stop words list provided by the `NLTK` library. To compare the performance of our model we chose two embedding-based baselines, which do not explicitly utilise global information such as TF-IDF: skim-gram Word2Vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017) embeddings. We found that for both 200-dimensional representations performed best. The embeddings were passed through a logistic regression classifier, following the methodology suggested by De-Arteaga et al. (2019), with L_2 regularization. The results are shown in table 1.

Model	**Val Acc**	**Val Loss**	**Test Acc**	**Test Loss**
Word2Vec	0.7614	0.8083	0.7634	0.7982
FastText	0.7606	0.8034	0.7624	0.7945
TextGCN	0.7823	0.7408	0.7792	0.7515

Table 1: Comparison of 200-dimensional Word2Vec and FastText embeddings with logistic regression and a TextGCN with one 200-dimensional hidden layer.

Initially, when trying to construct the graph using the entire dataset, a significant computational difficulty was encountered. We considered building multiple smaller graphs, however, the main advantage of using a GCN, i.e. being able to capture global relationships across the entire corpus, would have been lost. Instead, we decided to include in the graph only words that occurred at least 5 times in the corpus. After the graph was built following the methodology outlined in section 3, it was passed through the GCN. We experimented with no hidden layer, a single hidden layer and two hidden layers. We found that using just the information from the immediate neighbours, i.e. no hidden layer, was not sufficient for the model to learn the representation of the nodes, while using 2 hidden layers lead to a drop in validation and test accuracies. Hence, we found that using a single 200-dimensional hidden layer gave the highest validation and test accuracies of 0.7823 and 0.7792 respectively. Experimenting with the hyperparameters, we arrived at the following final setup: learning rate of 0.02 and droupout probability of 0.5. We found that using glorot weight initialisations performed best and Adam performed better than SGD. We trained the model for 200 epochs, allowing for early stopping.

We find that our proposed model outperforms both of the baseline models with pre-trained word embeddings. We propose two reasons for that: firstly, the graph captures both global word-word relations and document-word relations. Moreover, the way the information is propagated in the GCN model means that the new feature representation of a node is an aggregation of the representation of the node's features and its neighbour nodes. Hence, label information of document nodes is propagated to neighbour word nodes and further passed onto document and word nodes that are one hop away from the word nodes. In this way, the word nodes are able to relay document label information throughout the whole graph. Also, it is important to notice that for this particular task the word order is of little importance, but if it was, we would not expect a stand-alone GCN to perform well. Also, we do not make numerical comparisons in this section with the experimental results obtained by De-Arteaga et al. (2019) since we are using only about a quarter of the BiosBias dataset.

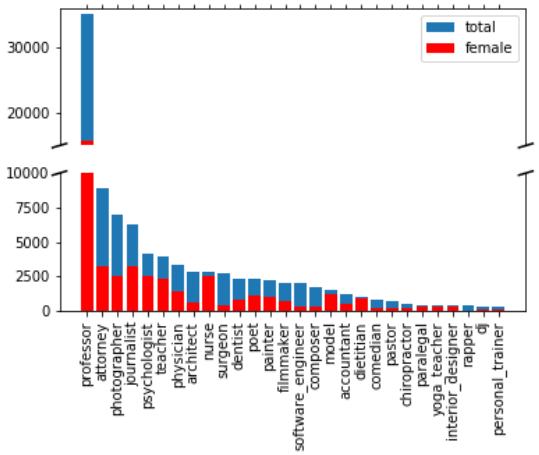


Figure 1: Distribution of the number of biographies in total and for female biographies only

4 Experiments

4.1 Empirical setup

In our experiments, we consider a subset of the BiosBias dataset assembled by De-Arteaga et al. (2019), which consists of biographies and corresponding occupation labels taken from the first sentence of each biography. The download and preprocessing was performed using modifications of the official GitHub repository for the re-creation of the dataset (Kalai, 2018). The subset comprises 97,798 biographies with 28 different occupation labels, for which the first sentence is excluded in the classification task. The original distribution of the number of biographies across the occupations is illustrated in figure 1 alongside the proportion of female biographies for the respective occupation. While *professor* is the prevalent occupation (35,136 biographies), the least frequent occupation is *personal trainer* (332 biographies). The proportion of female biographies varies from 12.3% (*rapper*) to 93.5% (*dietitian*).

To gain insights into the effects of varying sample sizes within the occupations and to draw conclusions for real world applications, we also consider a reduced dataset, where occupations are excluded for which less than 1,000 biographies are available. The resulting dataset consists of 93,536 biographies split into 19 occupations. The least frequent occupation changes to *dietitian* (1006 biographies) and the occupation with the lowest proportion of females to *surgeon* (14.2%).

For each of the two datasets, we examine two different scenarios: the classifier can use (1) the original biography including first name and explicit gender indicators, and (2) a scrubbed ver-

sion of the biography excluding first name and explicit gender indicators. The explicit gender indicators encompass the words *he, she, her, his, him, hers, himself, herself, mr, mrs, and ms*. For the experiments, the shuffled datasets are split into 70% training data and 30% test data. 10% of the training set is then used for validation. Comparing the proportion of female biographies across occupations for training and test sets, the proportion only varies by a maximum of 0.8% showing an equal representation in both training and test splits.

4.2 Metrics

We evaluate our experiments using the same metrics as De-Arteaga et al. (2019) to ensure comparability with their findings.

True positive rate gender gap: We quantify the gender bias in the test set by computing the true positive rate (TPR) gender gap, which is defined as the difference in TPRs between the binary genders g and $\sim g$ for every occupation y:

$$\text{TPR}_{g,y} = P(\hat{Y} = y | G = g, Y = y)$$
$$\text{Gap}_{g,y} = \text{TPR}_{g,y} - \text{TPR}_{\sim g,y}$$

where the random variable G denotes the binary gender of a biography's subject and the random variables $\hat{Y}$ and Y denote the predicted and target occupations respectively. The actual percentage of individuals with gender g in occupation y is defined as $\pi_{g,y} = P(G = g|Y = y)$. The correlation between the quantities $\text{Gap}_{g,y}$ and $\pi_{g,y} = P(G = g|Y = y)$ can give insights into the direction of their linear relationship.

Gender imbalance: Defining the initial gender imbalance of occupation y as $\frac{\pi_{g,y}}{\pi_{\sim g,y}}$, the gender g is underrepresented if $\frac{\pi_{g,y}}{\pi_{\sim g,y}} < 1$ or, equivalently, $\pi_{g,y} < 0.5$. If the underrepresented gender has a lower TPR than the overrepresented gender, e.g. $\text{Gap}_{g,y} < 0$ and g is underrepresented, De-Arteaga et al. (2019) show that the gender imbalance is compounded by a factor of $\frac{\text{TPR}_{g,y}}{\text{TPR}_{\sim g,y}}$. These quantities will be used to assess a potentially compounding effect of the classifier on the gender imbalance.

Counterfactuals: To examine what occupation the considered classifier would predict if a biography had used indicators corresponding to the other gender, we analyse the counterfactuals which are obtained by swapping gender indicators. Similar to De-Arteaga et al. (2019), we removed the first names in the test set and replaced explicit gender indicators by their complements, e.g. *he* by *she*. We then not only consider the percentage of predictions that change, but also examine pairs of occupations more closely. Therefore, we define a set of biographies $\mathbb{S}_{g,(y^1,y^2)}$ for which the occupation is incorrectly predicted as y^1 when using the original gender indicators, but correctly predicted as y^2 after swapping the gender indicators:

$$\mathbb{S}_{g,(y^1,y^2)} = \{x_i : \hat{y}_i = y^1, \hat{y}_i^{g \leftrightarrow \sim g} = y^2, y_i = y^2\}$$

where x_i denotes the i^{th} biography, y_i the corresponding occupation, $\hat{y}_i$ the predicted occupation for the original gender indicators, and $\hat{y}_i^{g \leftrightarrow \sim g}$ the predicted occupation after swapping gender indicators. When additionally defining the the total set of biographies for which the occupation y^2 is only correctly predicted after swapping the gender indicators as $\mathbb{S}_{g,y^2}$, the percentage of these biographies for which the label prediction changes from y^1 to y^2 after the gender indicators are swapped can be determined as:

$$\Pi_{g,(y^1,y^2)} = \frac{|\mathbb{S}_{g,(y^1,y^2)}|}{|\mathbb{S}_{g,y^2}|} \times 100\%.$$

5 Results and discussion

In this section, we will quantify and try to mitigate the gender bias when occupation classification is performed using a GCN. We will analyse the performance of the classifier using the metrics outlined in section 4.2 on the test splits of the two differently-sized datasets for two scenarios as described in section 4.1.

5.1 With explicit gender indicators

The following analyses consider the scenario in which first name and explicit gender indicators are available to the classifier. Figure 2 illustrates that almost all occupations in the original test set exhibit a gender imbalance, i.e. have different percentages of men and women. Only the occupations *journalist*, *poet* and *painter* are roughly balanced in both datasets. For occupations with an underrepresentation of women, i.e. $\pi_{female,y} < 0.5$, $\text{Gap}_{female,y}$ tends to be negative, whereas for occupations where $\pi_{female,y} > 0.5$, i.e. women are overrepresented, $\text{Gap}_{female,y} > 0$ holds with the exception of *yoga teacher*, which is not included in the reduced dataset. The converse relationship holds true for $\pi_{male,y}$ and $\text{Gap}_{male,y}$. Similar to De-Arteaga et al. (2019), we detect a positive correlation between the TPR gender gap

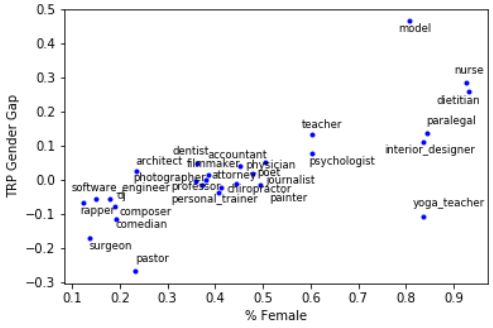


Figure 2: $\text{Gap}_{female,y}$ against $\pi_{female,y}$ with explicit gender indicators on the test split of the original dataset

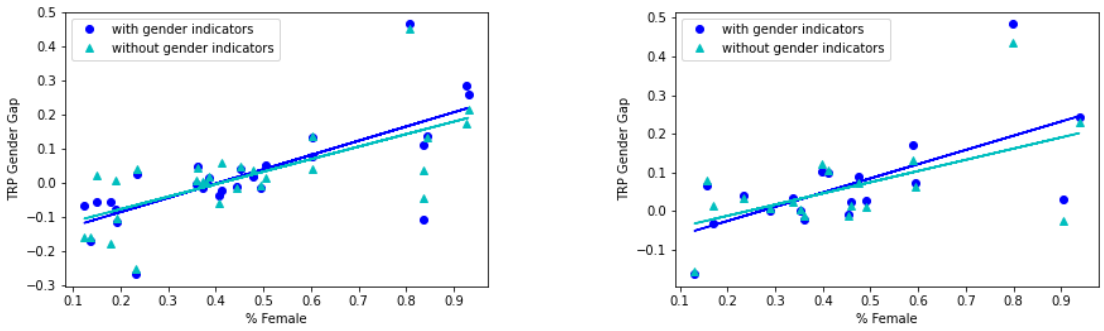
and the gender imbalance for an occupation y regardless of whether the original dataset or the reduced dataset is considered. To be more specific, the correlation between $\pi_{female,y}$ and $\text{Gap}_{female,y}$ on the original test split is 0.73. The correlation decreases to 0.66 when the reduced test split is examined, allowing the conclusion that having more data per occupation either mitigates the gender bias when classifying using a GCN or it excludes the occupations with the highest gender imbalances. However, comparing the correlation for the original test split to the results of De-Arteaga et al. (2019), the GCN reduces the correlation between the TPR gender gap and the gender imbalance considerably by at least 0.09 (compared to the model picking up the least bias, i.e. an RNN).

This observed correlation means that the employment of the classifier in a recruitment process for occupation y would result in the gender imbalance being compounded by the factor $\frac{\text{TPR}_{g,y}}{\text{TPR}_{\sim g,y}}$, where g denotes the underrepresented gender. Taking *composer* as an example, $\pi_{female,composer} < 0.5$ since the proportion of women in the original test split is 19%. Then, the $\text{TPR}_{female,composer}$ suggests that the classifier is able to predict a female composer in 76.2% of the cases if the biography indeed profiles a composer, while $\text{TPR}_{male,composer} = 0.838$ implies that the classifier correctly classifies the male composer in 83.8% of the cases. Therefore, $\text{Gap}_{female,composer} < 0$. The compounding factor of the classifier for this occupation is then 0.91 meaning that only 14.5% of the true positives are female when considering the original dataset. On the reduced dataset, the compounding factor increases to 0.976 meaning that starting from $\pi_{female,composer} = 0.171$, the proportion of women among the true positives is only decreasing to 16.7%. Thus, excluding occupations with few biographies reduces the bias that the classifier picks up for the occupation *composer*.

Similar to De-Arteaga et al. (2019), we analyse the effects of explicit gender indicators on the classifier's predictions by comparing the predictions on the two test splits as described before to predictions on two corresponding test splits, in which first names are removed and explicit gender indicators are swapped for their complementing version. By examining these counterfactuals, we would expect an unbiased classifier to predict the same occupation for a biography regardless of the gender. For the original test split, 5.5% of predictions change when the explicit gender indicators are swapped, while this proportion increases to 7.8% when the reduced test split is considered. So, there are a number of cases in which the GCN would predict a different occupation if the biography was referring to a man instead of a woman and vice versa. When examining $\Pi_{female,(y^1,y^2)}$ and $\Pi_{male,(y^1,y^2)}$, there are a number of (y^1, y^2) pairs for which $\Pi_{g,(y^1,y^2)} = 1$ implying that if y^2 is only correctly predicted after swapping explicit gender indicators, all of these biographies were previously falsely classified as having occupation y^1. On the original test split, the pairs for $\Pi_{female,(y^1,y^2)}$ include for instance (*composer*, *rapper*) and (*professor*, *accountant*), while (*poet*, *photographer*) and (*surgeon*, *dietitian*) are examples for $\Pi_{male,(y^1,y^2)}$. On the reduced test split, further examples comprise (*professor*, *physician*) for $\Pi_{female,(y^1,y^2)}$ and for (*photographer*, *model*) $\Pi_{male,(y^1,y^2)}$. The quoted cases encompass pairs in which the occupations exhibit gender imbalances, proving that the GCN still produces biased predictions. However, both datasets also include occupation pairs, for which $\Pi_{g,(y^1,y^2)} = 1$, that are the same for $\Pi_{female,(y^1,y^2)}$ and $\Pi_{male,(y^1,y^2)}$ such as (*photographer*, *painter*) and (*photographer*, *interior designer*) for the original test split and (*professor*, *journalist*) for the reduced test split. This implies that the classifier only confuses these occupations without a manifestation of bias. Such cases have not been reported for any of the models considered by De-Arteaga et al. (2019).

5.2 Without explicit gender indicators

In the second considered scenario, the classifier can only use scrubbed versions of the biographies



(a) Original dataset

(b) Reduced dataset

Figure 3: $\text{Gap}_{female,y}$ against $\pi_{female,y}$ with and without explicit gender indicators

with first names and all explicit gender indicators being removed. Assuming that men and women present themselves similarly in their biographies, the exclusion of explicit gender indicators should be enough to remove all gender information from the biographies. Then, we would expect the TPRs for men and women to be equal. In general, scrubbing explicit gender indicators shrinks the TPR gender gaps as can be seen in figure 3a for the original test split. However, $\text{Gap}_{female,y}$ remains large in absolute value for specific occupations such as *nurse* (0.454 instead of 0.467) or *pastor* (-0.25 instead of -0.266). For the reduced test split, figure 3b seems to reveal a greater reduction of the TPR gender gap when scrubbing explicit gender indicators, though this observation is likely to be driven by the exclusion of the occupations with less then 1,000 biographies. When comparing figure 3a and 3b, it is apparent that the excluded occupations tend to be those with high TPR gender gaps. Consulting the correlation coefficients between $\pi_{female,y}$ and $\text{Gap}_{female,y}$ presented in table 2 provides more evidence that scrubbing explicit gender indicators reduces the strength of the linear relationship between these two measures, implying that the GCN picks up less bias when explicit gender indicators are removed. Considering the original test set, the difference between the two scenarios is smaller than what De-Arteaga et al. (2019) report for their examined classifiers indicating that the GCN does not put as much emphasis on the explicit gender indicators as the approaches analysed by De-Arteaga et al. (2019). The correlation reduction effect is particularly large for the reduced test set, however, it has to be considered that occupations with larger TPR gender gaps are driving the results when included.

Scenario	**Original**	**Reduced**
w/ gender indicators	0.733	0.655
w/o gender indicators	0.689	0.562

Table 2: Correlation between $\pi_{female,y}$ and $\text{Gap}_{female,y}$

When examining the compounding effect of the GCN classifier on the gender imbalance, it can be noted that for 64.3% of the occupations in the original test split the compounding factor is pulled towards 1 after scrubbing explicit gender indicators from the biographies. Occupations for which $\text{Gap}_{female,y} > 0$ account for 35.7%, while the remaining 28.6% come from occupations for which $\text{Gap}_{female,y} < 0$. Thus, removing explicit gender indicators is more effective at mitigating gender bias for occupations in which women are overrepresented. Similar conclusions can be drawn when analysing the results for the reduced dataset: for 78.9% of the occupations, the compounding factor is pulled towards 1. 52.6% are contributed by occupations with $\text{Gap}_{female,y} > 0$, whereas 26.3% relate to occupations with $\text{Gap}_{female,y} < 0$.

5.3 Discussion

Our dataset is significantly reduced from its original size to allow tractable computational times. This is a common issue encountered when training large-scale GCNs, as the computational cost of SGD-based models scales exponentially with the number of GCN layers and the storage requirement for the whole graph and the embeddings of individual nodes is very large. This could be addressed through utilising the Cluster-GCN (Chiang et al., 2019), which accelerates training time and reduces storage requirements through

performing graph clustering, defining subgraphs and at each step in training sampling nodes from a block of nodes defined by the subgraph.

Moreover, all comparisons made with the results of De-Arteaga et al. (2019) rely on the assumption that the used subset is representative for the entire BiosBias dataset. The only available information about the composition of the BiosBias dataset are the ranks of occupations in terms of numbers of biographies per occupation (see De-Arteaga et al., 2019, p. 3, Figure 1). This allows us to calculate the Spearman's rank correlation coefficient ρ to evaluate how well the relationship between the occupations ranked by number of biographies of the BiosBias dataset and the subset can be described by a monotonic function. If all occupation ranks are equal, we would expect both datasets to roughly have the same structure and thus be described by a positive monotonic function, i.e. $\rho = 1$. We find that $\rho = 0.973$ and that this value is significantly different from zero on a 1% level. Therefore, we conclude that the subset is sufficiently representative for comparisons, although more information about the full BiosBias dataset is necessary for further validation.

To investigate the effects of the small dataset in more detail, a reduced dataset including only occupations with more than 1,000 biographies was examined. However, this approach mainly excluded occupations with high gender imbalances, which are of special interest when quantifying gender bias. This problem of few observations per occupation is arising in real world applications, where companies have limited access to data. Hence, it is necessary to further investigate optimal sample sizes to better understand the relationship between accuracy improvements due to more available data and the extent to which the bias is compounded for varying dataset sizes.

6 Conclusion

This paper extends the work of De-Arteaga et al. (2019) by quantifying and mitigating gender bias in occupation classification with GCNs. Our GCN, consisting of a single 200-dimensional hidden layer, outperforms both baselines models, which use Word2Vec and FastText embeddings, with respect to validation and test accuracies.

TPR gender gaps are present in our GCN predictions. However, we find the correlation between TPR gender gaps and gender imbalances to be considerably smaller than the values reported by De-Arteaga et al. (2019). Assuming that the BiosBias subset we use is representative of the entire dataset, the GCN picks up less bias than the semantic representations bag-of-words, word embeddings, and deep recurrent neural networks examined by De-Arteaga et al. (2019). Removing explicit gender indicators from the biographies reduces the correlation between the TPR gender gaps and the gender imbalances further when using a GCN as classifier. Likewise, the compounding effect of the GCN on the gender imbalance is reduced by scrubbing explicit gender indicators.

Excluding occupations with less than 1,000 occupations is found to reduce gender bias in terms of a smaller correlation between TPR gender gaps and gender imbalances, and a lower compounding effect of the GCN on the gender imbalance. However, these results need to be interpreted with care as the dataset reduction removed occupations with particularly high gender imbalances. These findings highlight the necessity of further research into how different dataset sizes affect the extent to which predictions are biased with important implications for automated recruitment processes.

While this project specifically explores gender bias, NLP-based models used in automated hiring can also develop other forms of bias, such as racial, religious or socio-economic bias. It is important that more research is performed into understanding how these biases arise and what can be done to prevent this. One of the benefits of using a GCN in this context is that there now exist frameworks such as the GNNExplainer (Ying et al., 2019) which attempt to dispel some of the ambiguity originating from the employment of deep learning approaches as 'black-box' models. The GNNExplainer generates a subgraph of neighbouring nodes and edges which were influential in the classification process and thus might be a starting point to find new bias mitigation strategies.

References

Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph convolutional encoders for syntax-aware neural machine translation. *arXiv preprint arXiv:1704.04675*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. 2019. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 257–266, New York, NY, USA. Association for Computing Machinery.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

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. *arXiv preprint arXiv:1903.03862*.

Adam Kalai. 2018. Biosbias. *GitHub repository*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907.

Yifu Li, Ran Jin, and Yuan Luo. 2018. Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks (Seg-GCRNs). *Journal of the American Medical Informatics Association*, 26(3):262–268.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Cite arxiv:1301.3781.

Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2019. Dating documents using graph convolution networks. *arXiv preprint arXiv:1902.00175*.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2019. A comprehensive survey on graph neural networks. *CoRR*, abs/1901.00596.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2018. Graph convolutional networks for text classification. *CoRR*, abs/1809.05679.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

A Appendices

A.1 Summary statistics for reduced dataset

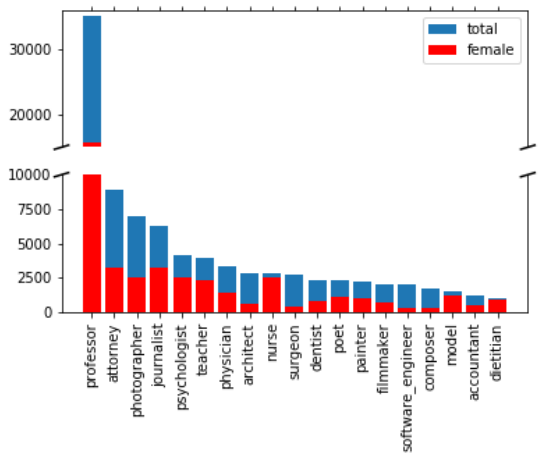


Figure 4: Distribution of the number of biographies across the 19 occupations of the reduced dataset in total and for female biographies only

A.2 TPR gender gaps for scrubbed original dataset

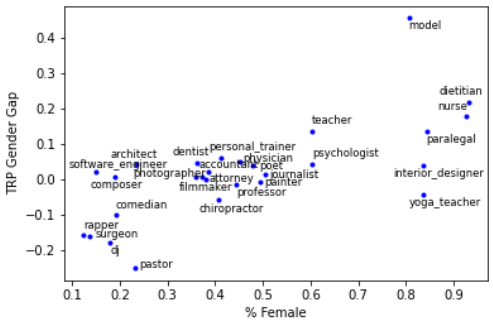


Figure 5: Plot of $\text{Gap}_{female,y}$ against $\pi_{female,y}$ with explicit gender indicators on the scrubbed test split of the original dataset

A.3 TPR gender gaps for reduced dataset

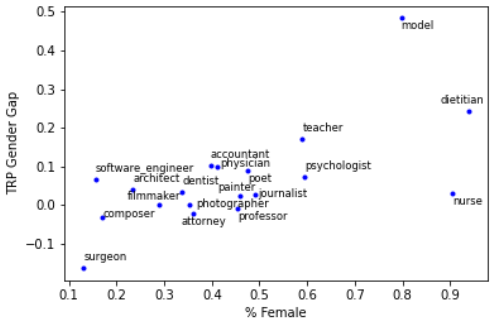


Figure 6: Plot of $\text{Gap}_{female,y}$ against $\pi_{female,y}$ with explicit gender indicators on the test split of the reduced dataset

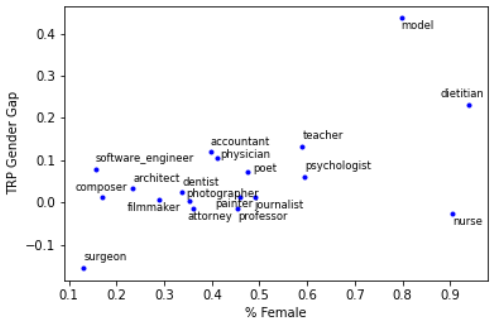


Figure 7: Plot of $\text{Gap}_{female,y}$ against $\pi_{female,y}$ with explicit gender indicators on the scrubbed test split of the reduced dataset

A.4 Visualisations of word and document embeddings learned by TextGCN

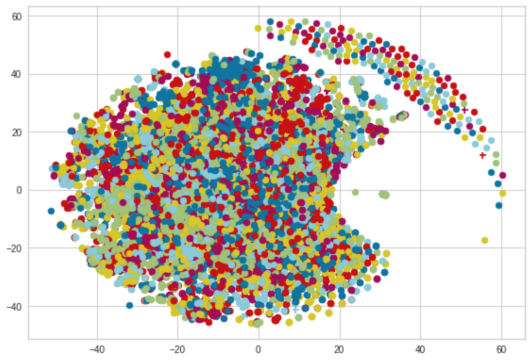


Figure 8: Visualisation of word embeddings learned by the output layer of the GCN, using the t-SNE tool.

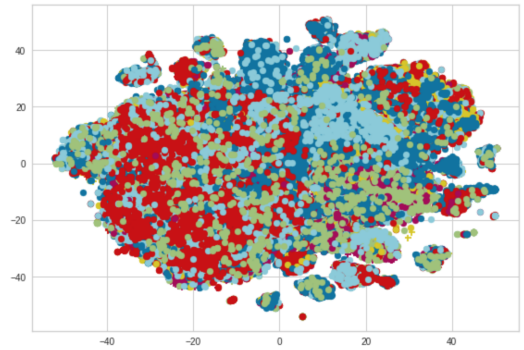


Figure 9: Visualisation of document embeddings learned by the output layer of the GCN, using the t-SNE tool.