Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Econometrics and Management Science

Business Analytics and Quantitative Marketing

Towards More Equitable NLP: Investigating Multi-Sensitive Attribute Debiasing Methods for Contextualised Models

---

Author: Meike Pluym (466476)

Supervisor: dr. Hakan Akyuz

Second assessor: dr. Flavius Frasincar

April 3, 2023

Abstract

NLP-based applications have grown to become an integral part of our daily tasks. Such NLP models, however, through capturing human semantics, inevitably encode harmful social biases as well. Research into bias mitigation in the state-of-the-art textual inference models has, therefore, become necessary. This paper focuses on the attenuation of social biases from contextualised NLP models, specifically researching the possibilities for debiasing with respect to multiple sensitive attributes simultaneously. To this end, the paper investigates three methods for Multi-Sensitive Attribute (MSA) debiasing of such models. The first method (Loss) penalises stereotypical associations in the training objective of the NLP models, while the second (INLP) and the third (R-LACE) produce projection matrices which allow the projection of embedding structures onto the orthogonal complement of the bias subspace in these embeddings. The results show that all methods are capable of MSA bias reduction in contextualised NLP models. Ultimately, the Loss method indicates to be the most promising with respect to the inevitable accuracy vs. fairness trade-off in debiasing the models. Though all methods decrease harmful stereotypical associations in the models, the Loss method does so whilst retaining an appropriate level of semantic information inherent in the NLP model.

Contents

1 Introduction

Daily tasks, nowadays, are more and more facilitated by the use of Natural Language Processing (NLP). From ubiquitous applications, such as search engines, e-mail filters, and predictive text, to large company uses, such as modeling customer intent or automating résumé screening (Black and van Esch, 2020). A recent significant progression in our everyday NLP use, is the launch of 'ChatGPT' (OpenAI, 2021), which has made the advanced GPT-3 NLP model readily accessible to the public in the form of a human-like response machine. NLP is a sub-field of Machine Learning (ML) which considers natural language, by constructing numeric representations of text, based on co-occurrence statistics (Caliskan et al., 2022). These numeric representations are generally referred to as 'word embeddings'. In order to construct word embeddings, we train algorithms on large, authentic text corpora, such as sets of Amazon reviews, Wikipedia links or Google News feeds.

Research has shown that, whilst this method for constructing word embeddings captures human semantics well, it also inevitably encodes types of human social biases. NLP models trained on authentic, human-produced datasets necessarily contain associations inherent in human thought, including "unfair" bias (Bolukbasi et al., 2016; Garg et al., 2018; Dev et al., 2020; Kaneko and Bollegala, 2021; Caliskan et al., 2022). Caliskan et al. (2017) are among the researchers who investigated the presence of human bias in word embeddings. The researchers show the presence of historic biases in text corpora. They also emphasise that, whilst these biases can be morally neutral as towards insects or plants, they can also be problematic as towards race or gender.

The encoded problematic biases in word embeddings can lead to incorrect or unfair decisions, as they can be discriminatory based on the associated stereotypes. Bolukbasi et al. (2016), for example, conclude in their research that embeddings which are trained on Google News articles exhibit typical female/male gender stereotypes greatly. In addition, Berkeley's Computation and Language Lab recently shared ChatGPT responses to requests for writing a Python program which could determine "which air travelers present a security risk." As a response, OpenAI's ChatGPT model proposed the calculation of a traveler's 'risk score' which would increase if the person is Syrian, Iraqi, or Afghan (Biddle, 2022). Using word embeddings in decision making processes, therefore, can cause discriminatory practices as the decisions might be influenced by stereotypes of sensitive attributes, such as gender, nationality, religion, and race. As discrimination is increasingly considered intolerable from social, ethical, and even legal perspectives, the need for research into bias mitigation in NLP models grows as well.

Most research into bias mitigation of word embeddings focuses on static (non-contextualised) word embeddings. These embeddings represent a word by a single vector in different contexts. Contextualised embeddings, however, provide context dependent embeddings for different instances of words. Allowing the embedding to be context-dependent has significantly improved performance in various NLP tasks (e.g. Devlin et al., 2018). Therefore, these contextualised embeddings have been established as the state-of-the-art 'type' word embedding for text representation (Kaneko and Bollegala, 2021). Research has shown that, unfortu-

nately, these contextualised embeddings encode unfair stereotypical bias, similarly to static embeddings (e.g. Zhao et al., 2019). Kaneko and Bollegala (2021) express in their research that, whilst various studies have reported the existence of unfair bias in contextualised embeddings, methods for mitigating this bias are fairly unexplored.

In addition to this open problem in research, there is also the problem of debiasing embeddings with respect to multiple sensitive attributes simultaneously. Current methods show promising results in the attenuation of a single bias dimension, such as gender or race or nationality (e.g. Lahoti et al., 2019; Zehlike et al., 2017). However, in reality, multiple bias dimensions can be present in a dataset. This research, therefore, investigates methods which can mitigate bias with respect to various sensitive attributes simultaneously. The investigative goal of this research can be summarised into the following main research question:

How can we attenuate bias in textual analysis whilst retaining an appropriate level of semantic information?

Specifically, this research investigates the main research question with respect to Multi-Sensitive Attribute (MSA) debiasing, in combination with the analysis of mitigating bias in contextualised embeddings. To this end, two sub-questions are proposed which, through analysis, give us novel insights into the attenuation of bias in textual analysis, as described in the main research question. The sub-questions are

1. *How can we accurately decrease bias in contextualised word embeddings?*
2. *Can we attenuate bias related to several sensitive attributes simultaneously?*

The ability to take into account multiple bias directions at the same time, provides a more realistic form of bias mitigation in real-life datasets. Furthermore, focusing on contextualised embeddings gives us a chance to attack the fairness problem for the state-of-the-art NLP systems. Overall, therefore, the combination of these two focus points can provide a novel, pragmatic approach for mitigating bias in currently used NLP models and applications.

In order to answer the research questions, three different techniques for bias attenuation in contextual NLP structures are proposed. The first method uses a weighted loss function which, while minimising the decrease in accuracy of the model, minimises the 'similarity' between embeddings of sensitive attribute words and their related stereotypes. This loss function is implemented in the training objective of the NLP model and uses attribute and (stereotyped) target words related to several sensitive attributes to identify the bias structures in the model.

The second method used to decrease bias in the NLP models, is an iterative approach which constructs a more complete version of a projection matrix in each step of the algorithm. This projection matrix, eventually, captures the correlations between target words and their corresponding sensitive attribute words, in the form of bias directions in the word embedding. This projection matrix, then, projects the model output onto the orthogonal complement of the bias subspace, where the bias subspace is a linear subspace of the embedding

which is spanned by these bias directions. Effectively, this method produces alternative word embeddings which include as little of the underlying biases in the model as possible.

The third debiasing method, similar to the second, uses a projection matrix to remove stereotypical associations from the word embeddings. This method, however, constructs an elaborate minimax game, which explicitly attempts to remove as much of the bias, whilst removing as little of the semantic influences from the word embeddings as possible.

For each of the methods, this research investigates the effectiveness of removing implicit biases in the word embeddings. In addition, there is a focus on identifying the debiasing method which, while removing the bias structures, retains most of the semantic information of the NLP models. This investigation is performed with respect to several NLP models, in order to analyse the level of universality of the debiasing method with regards to different models.

Ultimately, the results show that each method is successful in attenuating MSA bias from the various NLP models. Each debiasing method manages to increase the equitability level of the models. In addition, the loss function method manages to retain a sufficient level of semantic information of the models. These results indicate the such method is able to provide a novel and pragmatic approach for the reduction of MSA bias in current state-of-the-art NLP applications.

The research paper is structured as follows. Sec. 2 describes existing research into the topic of bias reduction in machine learning. This section will highlight the extent to which bias reduction has been investigated previously, and how this research can provide new insights. Next, Sec. 3 provides an overview of the datasets and data generation approaches used in this research. Sec. 4 gives a detailed explanation of all methods used to achieve the results and answer the research questions. After, Sec. 5 provides an analysis of the results obtained by applying the various debiasing methods. Finally, Sec.6 provides the summary of the findings of the paper, as well as some areas for future research.

2 Literature Review

This section provides an overview of previous research into the field of bias reduction in NLP models. In order to give a complete outline of previous research, first the importance of NLP models and their various areas of everyday use are described. After, the existence of sensitive bias in these models is elaborated upon. Furthermore, the first and most prevalent efforts towards mitigating such bias are discussed. Finally, Sec. 2.3 and Sec. 2.4 describe research into mitigating bias in static versus contextualised embeddings, and on Single-Sensitive Attribute (SSA) bias versus MSA bias attenuation.

2.1 NLP Models and the Bias Present Them

The growing amount of research into NLP models and their applications ensures that NLP, nowadays, is commonly present in many downstream applications, such as speech recognition, text translation, search

completion, and information extraction (Nielsen, 2021). In everyday life, we, as individuals, and businesses (unconsciously) make use of the complex algorithms behind NLP applications. It is no surprise, therefore, that research into improving the performance of such word embedding models is popular. Bowman et al. (2015) show that it is possible for fixed-length representations of sentences to 'understand' logical deduction and logical semantics. Nalisnick et al. (2016), for example, improve the performance of Word2Vec embeddings for the ranking of documents with respect to specific queries. Furthermore, researchers such as Devlin et al. (2018) go further into showing that context-dependent embeddings can achieve significantly better performance in various NLP tasks compared to context-unaware embeddings. There is a myriad more examples of research aiming to improve embeddings or provide alternative models in order to increase their performance. However, relative to the amount of research into improving the performance of NLP models, there are few research papers that take into account any form of inherent and unfair bias in the model, although research has shown that such bias can be problematic.

Caliskan et al. (2017) show that text corpora contain recoverable and accurate imprints of human historic biases. They emphasise that, whilst some of these biases may be harmless, there are very problematic biases present as well. Garg et al. (2018), in addition, demonstrate that word embeddings can be used to quantify historical trends and social change. Their findings include the ability to use word embeddings to analyse the evolving gender and ethnic stereotypes during the 20th and 21st centuries, using text data of the past 100 years. Bolukbasi et al. (2016) show that word embeddings trained on Google News articles, a corpus which we would expect to contain little to no problematic bias, exhibit a great amount of gender stereotypes. Angwin et al. (2016) show the existence of bias in software used to predict future criminals. They find that such software is biased against black people. Lahoti et al. (2019) show that *www.xing.com* (a German job searching platform) ranked less qualified male candidates higher than more qualified female candidates. Kamiran et al. (2012) evaluate discrimination classification in real world datasets. They show that the bias found in 'artificial' environments (i.e. academic research data sets) is also present in real life. We can conclude that NLP models can propagate certain 'unfair' human biases found in text and image corpora. NLP models which contain such bias in their trained word embeddings, risk the introduction of the bias into real-world systems. In the past, this has already happened many times, even in the NLP applications of big, influential companies. For example, Google's facial recognition algorithm that labelled black people as gorillas (Guynn, 2015) and Amazon's résumé recommendation system which penalised the word "woman" in résumés (Dastin, 2018).

2.2 Mitigating Bias in Word Embeddings

Although the amount of research into bias mitigation is small relative to the amount of research into constructing word embeddings, some methods for bias mitigation have been investigated extensively. Generally, bias mitigation techniques can be categorized based on the 'stage' of the model in which it is deployed. In this sense, we can distinguish pre-processing algorithms, in-processing algorithms, and post-processing algorithms

(Zuo et al., 2018).

Pre-processing techniques often attempt to cancel out the disproportionate amount of references towards one 'side' of a stereotype (e.g. gender stereotypes) by augmenting the original data set. Zhao et al. (2018a), for example, create an augmented data set which is identical to the original, but biased towards the opposite gender. They essentially do a 'gender-swap' and, then, train their model on the union of the original and the swapped data set. Another approach for constructing an augmented data set is referred to as 'gender tagging'. This approach requires the addition of a tag, indicating the gender of the source, to the beginning of every data point and including this tag in the analysis (Sun et al., 2019). Generally, such pre-processing data augmentation methods are criticised for being very computationally heavy, since they require the amount of data in the data set to be, at least, doubled, without adding sentimental information (Dev and Phillips, 2019; Sun et al., 2019).

Post-processing techniques focus on the output labels. Such techniques generally take a subset of samples and change their predicted labels as to 'enforce' a degree of fairness (Lohia et al., 2019). Hardt et al. (2016), for example, propose the Equalized Odds Post-Processing method which changes output labels to create equal odds over some protected attribute. Zhao et al. (2018b) employ a retraining step on GloVe embeddings in order to mitigate bias. Lohia et al. (2019) train a classifier for the identification of individual unfairness on their output data. Samples which are classified to be likely to contain individual bias are then considered for a change of output label. These proposed methods provide promising results. The problem, however, is that such methods, again, have the danger of being computationally very expensive. A retraining step can add a large amount of computation time which can become a problem, especially given that these types of models are generally inherently expensive to train (Dev and Phillips, 2019).

All in all, it could prove beneficial to focus on less computationally heavy modifications to mitigate bias, which are not based on re-training a model or doubling the amount of data in the data set.

In-processing methods generally 'alter' the optimization step in the training process of algorithms. Therefore, they do not require the adjustment of the data set, or the addition of a retraining step. In terms of bias mitigation, several modern, in-processing, mitigation techniques find and utilize the bias direction in word embeddings to reduce unfair bias. Finding this bias direction can be accomplished by various feature subspace determination methods, such as Principle Component Analysis (PCA) (Bolukbasi et al., 2016; Manzini et al., 2019; Dev et al., 2020). Knowing the bias direction, one can investigate the cosine similarity between certain words, names, verbs, or adjectives and the bias direction vector. The bias direction and cosine similarity can then be used to remove or disentangle the bias from the unrelated words. Bolukbasi et al. (2016), for example, project all word embeddings in their data set orthogonally to the identified gender bias direction, essentially 'removing' the influence of the bias direction from the embedding. By doing so, they find a significant decrease in gender bias. Dev et al. (2021) use a more subtle 'correction' method. They use a balanced approach of mitigation. Their approach focuses more on 'disentanglement' from the bias than on removal, with the aim of retaining a higher degree of sentiment whilst still mitigating unwanted associations.

2.3 Mitigating Bias in Contextualised Embeddings

Past research has mainly investigated bias mitigation with respect to non-contextualised (static) word embeddings (e.g. Schmidt, 2015; Bolukbasi et al., 2016; Manzini et al., 2019; Dev et al., 2021). There exists less research on the mitigation of bias in contextualised word embeddings. In regards to the existing research into the mitigation of bias in such embeddings, we can distinguish two 'forms' of bias mitigation approaches.

The first form focuses on the optimization criteria in the training phase of the contextual model. Kaneko and Bollegala (2021) propose a fine-tuning method for the debiasing of contextualised embeddings. They introduce a loss function into the hidden layers of the models. This loss function, if minimized, forces the hidden states to be orthogonal to some protected attribute. The researchers show that this method works well for bias mitigation. This method, however, still uses non-contextualised representations of the protected attribute, meaning that the loss function is still restricted to some form of non-contextualization.

For the second form, researchers, such as Dev et al. (2020), use the original techniques used for the debiasing of static embeddings to debias the 'static part' of the contextualised embedding. The static part of a contextualised embedding is the part of the embedding which is identical for different instances of the same word. In the BERT model, for example, this would be the sub-word (token) embeddings (Dhami, 2020). Dev et al. (2020), show that the method of removing bias from the static part of a contextualised word embedding is effective in reducing the bias in the overall contextual model.

2.4 Debiasing Multiple Bias Directions Simultaneously

Existing research into bias mitigation generally contains a common deficit. Almost all research into debiasing focuses on mitigating merely one bias direction. Focusing on one bias direction necessarily means that only one sensitive attribute can be protected from unfair bias. In other words, modern research mainly investigates the attenuation of either gender bias, or racial bias, or religious bias, etc. (e.g. Bolukbasi et al., 2016; Manzini et al., 2019).

Popović et al. (2020) were the first to propose joint multiclass debiasing techniques. The techniques they use are post-processing methods which aim to minimize some bias level criterion, whilst maintaining word relationships. The researchers show that their method is a good starting point for ongoing effort towards debiasing with respect to various sensitive attributes simultaneously, and by doing so, providing more unbiased neural representations of textual data.

3 Data

In this section, the datasets used in this research are identified and explained. In the research, two separate datasets are used, a set used for training and a set for evaluating the performance of the models. In addition, various word lists are used to extract specific text portions from the bigger (train) dataset. Sec. 3.1 outlines

the training dataset, the word lists, and the process used to extract the text portions. After, Sec. 3.2 gives an overview of the second dataset, used for performance evaluation.

3.1 Training Data

The dataset used for constructing the several debiasing techniques in the NLP models is the News-Commentary-V15 corpus (EMNLP, 2021). This corpus is a collection of news articles and opinion pieces from various sources. The articles included in the corpus are collected from newspapers, magazines and online news websites from many different regions of the world. All articles included were published between 2007 and 2016. Furthermore, the topics included in the articles cover a great range of subjects. The dataset is widely used in NLP research (e.g. Kaneko and Bollegala, 2021), because of its diverse and large nature.

In order to construct the debiasing techniques used in this paper, the training data should be comprised of a set of texts containing words which indicate the sensitive attributes and stereotype words associated with these attributes. In order to establish such training data, word lists are used to extract text fractions from the News-Commentary-V15 corpus containing such sensitive and stereotype words. Using, for example, a collection of gendered words (e.g. "he", "she", "his", "hers") and a collection of associated stereotypes (e.g. occupations and personality traits), we can extract portions of text from the corpus which contain one or more of these words. In this paper, words related to sensitive attributes are referred to as 'attribute words'. Words related to stereotypes of these sensitive attributes are referred to as '(neutral) target words'. The attribute and target word lists used in this research are comprised of lists from various published researches. With regard to the gendered attribute words, word lists published by Zhao et al. (2018b) are used. The corresponding target word list used, is that published by Kaneko and Bollegala (2019). With regards to the attribute and target words corresponding to other sensitive attributes, such as race and religion, we use a selection of word lists published by Dev et al. (2020). Appendix B.1 and Appendix B.2 give summarised overviews of the attribute and target words used in this research.

The attribute and target word lists are, thus, used to extract training data sentences containing them from the News-Commentary-V15 corpus. These training data sentences can then be used to identify bias directions in the NLP models and to train them to attenuate these biases.

3.2 Evaluation Data

In addition to the dataset used for debiasing the NLP models, another dataset is utilised to evaluate the performance of the final models. To do so, we use the *StereoSet* dataset (Nadeem et al., 2021), which is constructed to measure stereotypical biases inherent in NLP models as well as to evaluate the language modelling ability of the models.

The dataset includes thousands of associated sentence triplets. Each sentence includes some protagonist from a certain demographic and a description, action or event related to the protagonist. The demographics included in the dataset are gender, occupations, race, and religion. For the sentence triplets, the protagonist

remains the same. However, the context, or association of the sentence is different for the three sentences. Each instantiation of the sentence corresponds to one of the following three; a stereotypical association, an anti-stereotypical association, or a meaningless association (Nadeem et al., 2021). The stereotypical and anti-stereotypical instantiations of the sentence are used to measure the bias in the model. The meaningless association provides a benchmark for the language modeling ability of the model.

An example of such sentence triplets is:

Stereotype:	Our housekeeper is a Mexican.
Anti-stereotype:	Our housekeeper is an American.
Meaningless:	Our housekeeper is a fish.

Using such sets of sentences, we can evaluate the model's ability to rank meaningful contexts higher than meaningless ones. Based upon the concept of cognitive associations (Devine, 1989), both the stereotypical and the anti-stereotypical context of the sentence should rank higher than the meaningless context. Given the prevalence of cognitive associations, which refer to mental connections between concepts formed through experience, socialisation and cultural norms, both stereotypical and anti-stereotypical associations are, in practice, more probable to be made than meaningless ones. In addition, we can evaluate the extent to which the model is biased. An unbiased model should not rank the stereotypical context over the anti-stereotypical one (or vice versa), as an unbiased entity would view them as equally probable.

4 Methodology

In the following subsections, the methods used for answering the research questions are proposed. First, the various debiasing methods are explained. After, a description of how and where these methods are implemented in the NLP models used for the evaluation of the methods is given. Lastly, the performance measures used to achieve a final conclusion are described. The mathematical definitions of this section are used consistently across the subsections, and an overview of all definitions can be found in Appendix A.

4.1 Weighted Loss Function

The first debiasing method is based on work by Kaneko and Bollegala (2021). This method proposed by the researchers is extended in order to achieve a reduction of bias in several sensitive attributes simultaneously. The method is based around defining an extensive weighted loss function which can be embedded within the optimisation process of contextualised NLP models. By minimising the constructed loss function, MSA bias in the model's embeddings is reduced.

The weighted loss function consists of two formal requirements and uses the following definitions. A word (or token) is represented by $w \in \mathbb{R}^d$, where d is the dimension of the word embedding. As stated in Sec. 3.1, we can define two specific 'types' of words, sensitive attribute words and (neutral) target words. The set of

words belonging to sensitive attribute c is denoted by Λ_a^c. Here, c stands for a type of sensitive attribute, such as gender, race and religion. Similarly, the set of target words related to attribute c is given by Λ_t^c. Furthermore, the set of sentences containing word (or token) w is denoted by $\Omega(w)$. Using these definitions, $A_c = \bigcup_{w \in \Lambda_a^c} \Omega(w)$ and $T_c = \bigcup_{w \in \Lambda_t^c} \Omega(w)$ denote the sets of sentences containing the attribute and target words of sensitive attribute c, respectively. Using these definitions, the formal objective of this debiasing method is to alter the embedding model in order to retain semantic information with respect to $w \in \Lambda_a^c$, while eliminating implicit biases with respect to $w \in \Lambda_t^c$, for $c = 1, ..., C$.

With regards to the NLP model structures used in this research, E denotes the contextualised NLP embedding model in question. The pre-trained parameters of model E are, then, denoted by θ_e. Using x to denote the input sentence, we can then define $E_i(w, x; \theta_e)$ to be the embedding of token w in sentence x in the i-th hidden layer of E, where $i = 1, ..., N$. Lastly, for the loss function, we require a representation of the non-contextualised embedding of w in the i-th layer. Following Bommasani et al. (2020), this non-contextualised embedding of w is defined as the average of the contextualised embeddings of w in layer i of embedding model E over the whole set $\Omega(w)$. Formally, the non-contextualised embedding $\nu_i(w)$ is, thus, defined as

$$\nu_i(w) = \frac{1}{|\Omega(w)|} \sum_{x \in \Omega(w)} E_i(w, x; \theta_e), \tag{1}$$

where $|\Omega(w)|$ is the cardinality of set $\Omega(w)$.

Using the definitions given above, we can now construct the formal requirements which construct the final loss function of this debiasing method. The first requirement is that the target words must contain as little information related to the sensitive attribute words as possible. Formally, we can define this to require the inner product of the non-contextualised attribute word embeddings $\nu_i(a)$, where $a \in \Lambda_a^c$, and the contextualised target word embeddings $E_i(t, x; \theta_e)$, where $t \in \Lambda_t^c$, to be minimised for $c = 1, ..., C$. This first part of the final loss function can, thus, be defined by the requirement.

$$\min_{\theta_e} L_i = \sum_{c=1}^{C} \sum_{t \in \Lambda_t^c} \sum_{x \in \Omega(t)} \sum_{a \in \Lambda_a^c} \left(\nu_i(a)^\intercal E_i(t, x; \theta_e)\right)^2 \tag{2}$$

Minimising L_i forces the hidden states of E to be as close to orthogonal to the each of the sensitive attributes as possible.

The second requirement is that the debiased embeddings should preserve the useful semantic information in the model as much as possible. The debiased embeddings $E_i(w, x; \theta_e)$ should not be too far removed from the original, pre-debiasing, embeddings of the model, defined by $E_i(w, x; \theta_{pre})$. We can define this requirement as the minimisation of the euclidean distance between the debiased embeddings and the original embeddings. This second requirement can be viewed as the regularisation term of the overall loss function and is defined as follows.

$$\min_{\theta_e} L_{reg} = \sum_{c=1}^{C} \sum_{x \in A_c} \sum_{w \in x} \sum_{i=1}^{N} ||E_i(w, x; \theta_e) - E_i(w, x; \theta_{pre})||^2 \tag{3}$$

Above requirements, ultimately, construct an overall objective function, which is calculated using a weighted sum of the two loss functions. The weights in this objective function influence the accuracy versus fairness trade-off. The final loss function is

$$L = \alpha L_i + \beta L_{reg}, \tag{4}$$

for hidden layer i, where $\alpha, \beta \in [0, 1]$, such that $\alpha + \beta = 1$. In this research, following the investigation and results obtained by Kaneko and Bollegala (2021), we set $\alpha = 0.2$ and we implement the debiasing objective function in all hidden layers of the NLP model. Therefore, evaluating the overall objective function calls for the summation of the loss function for all $i = 1, ..., N$, resulting in the loss function

$$L = \sum_{i=1}^{N} 0.2L_i + 0.8L_{reg}. \tag{5}$$

Ultimately, loss function 5 is added to the original training objective of the various NLP models as a penalty. This way, the model penalises stereotypical associations, whilst, simultaneously, penalising a big variation from the original embeddings in the training process.

4.2 Iterative Null-Space Projection

The second debiasing method uses ideas proposed by Ravfogel et al. (2020). An adaptation of the Iterative Null-space Projection (INLP) method, proposed by these researchers, is used in a contextualised and MSA framework. For this approach, we use an iterative algorithm to capture and attenuate the bias in the last hidden layer of the embedding models.

The algorithm used in this debiasing method constructs a projection matrix which, if multiplied by the original hidden layers of the NLP models, produces an alternative representation of the word embedding vectors. This debiasing method, therefore, is centered around producing a projection matrix which will capture as much of the bias in the models as possible. The general approach for doing so requires the identification of the influence of the attribute words on the implicit biases in the target words, in the form of bias subspaces. The final alternative representations of the word embeddings are constructed by removing the bias subspaces from the original embedding vectors, using projection matrices. This way, the harmful stereotypes are removed from the original embeddings.

The algorithm which produces the projection matrices uses similar definitions as used in Sec. 4.1, complemented by the following definitions. We define W_a^c to be the set of embeddings, $E_i(w, x; \theta_e)$, of all $w \in \Lambda_a^c$. Respectively, W_t^c is defined to be the set of embeddings of all $w \in \Lambda_t^c$. In these terms, then, we need to find the bias subspace which contains the influence of W_t^c on W_a^c. This is done by training a Support Vector Machine (SVM), which predicts W_a^c from W_t^c. Training the SVM results in matrix M_0^c, which parameterises the model. This parameter matrix can be viewed as the (first version of the) bias subspace of sensitive attribute c.

Ultimately, we wish to acquire alternative vectors for W_t^c, such that M^c has no effect on the embedding vectors of $w \in \Lambda_t^c$. Formally, we wish to construct a projection matrix P^c, such that $M^c(P^c W_t^c) = 0$.

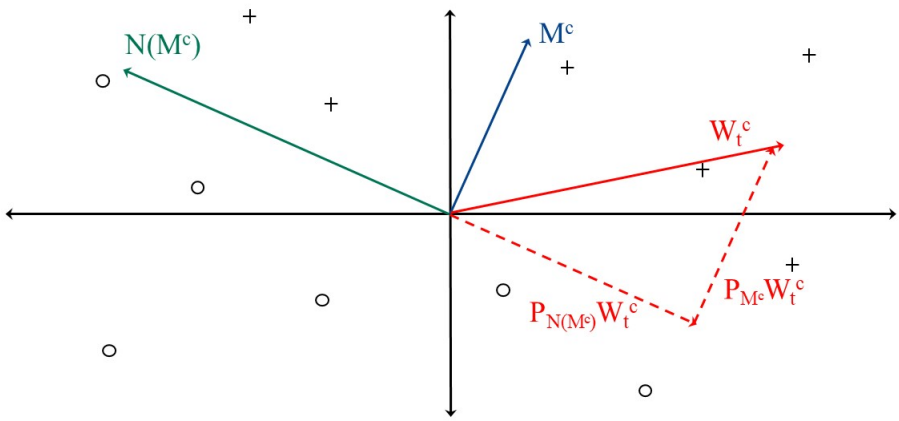

Figure 1. Simple (2D) representation of null-space projection

To construct such a projection matrix, we use the null-space of matrix M_0^c. The null-space of this matrix, $N(M_0^c)$, identifies the direction which is orthogonal to M_0^c, the influence of the sensitive attributes on the target words. Projecting W_t^c onto $N(M_0^c)$, therefore, zeroes the components of W_t^c which are influenced by M_0^c. Thus, using null-space $N(M_0^c)$, we can construct a projection matrix $P_{N(M_0^c)}$ with the property

$$M_0^c(P_{N(M_0^c)} W_t^c) = 0, \tag{6}$$

as represented in Figure 1. The construction of this initial projection matrix is done using the basis vectors of the orthogonal complement to the bias subspace, $N(M_0^c)$. The basis vectors, B_0^c, of $N(M_0^c)$ represent the subspace of the data which is orthogonal to the bias directions and, therefore, provide a projection, $P_{N(M_0^c)} = B_0^c B_0^{c\intercal}$, which removes that bias.

Having constructed projection matrix $P_{N(M_0^c)}$, we construct the first version of the alternative word embeddings of the target words, namely

$$Z_0^c = P_{N(M_0^c)} W_t^c. \tag{7}$$

Z_0^c, then, contains the embeddings of W_t^c from which the influence of M_0^c has been removed. In order to remove more and more of the influence of W_a^c on the attribute words, we repeat this process, replacing W_t^c by Z^c, until Z^c converges (or until a previously specified maximum number of iterations). An overview of the algorithm is given in Algorithm 1.

Algorithm 1 Algorithm for Iterative Null-space Projection

Input: Original word embeddings W_t^c, Initial bias subspace matrix M_0^c, Initialisation of the final projection matrix $P^c = I$, Some convergence threshold t, Maximum number of iterations F

1. Obtain the basis vectors B_0^c of M_0^c and compute the null-space projection matrix $P_{N(M_0^c)} = B_0^c B_0^{c\intercal}$

2. Initialise the debiased word embeddings $Z_0^c = P_{N(M_0^c)} W_t^c$

3. While $Z_f^c - Z_{f-1}^c > t$ **and** $f < F$ for $f = 0, ..., F$ **do**:

- **a.** Train the SVM on Z_f^c and W_a^c, obtain the bias subspace: M_f^c
- **b.** Compute the null-space $N(M_f^c)$ and use its basis vectors, B_f^c, to compute the new null-space projection matrix: $P_{N(M_f^c)} = B_f^c B_f^{c\intercal}$
- **c.** Compute the new debiased word embeddings: $Z_f^c = P_{N(M_f^c)} Z_{f-1}^c$
- **d.** Update the final projection matrix: $P^c = P_{N(M_f^c)} P^c$

4. Return projection matrix P^c

In every iteration of the algorithm, the influence of the attribute word embeddings on the (alternative representation of) the target word embeddings is reduced, removing more and more of the bias related to sensitive attribute c in the word embeddings with each step of the algorithm.

4.3 Relaxed Linear Adversarial Concept Erasure

The last debiasing method investigated in this paper is that of Linear Adversarial Concept Erasure (LACE) (Ravfogel et al., 2022). This method is similar to the INLP method as described in Sec. 4.2, in that it is also based on obtaining a projection matrix which acts as an adversary to the bias subspace in the model embeddings. The difference is in the construction of the projection matrix. The INLP method, as described above, involves the projection of the original data onto the null space of the matrix which captures the relationship between W_t^c and W_a^c. This involves the elimination of the basis of the identified parametarisation matrix (M^c) of a linear classification model (such as SVM), from W_t^c. The LACE method, on the other hand, involves the identification and elimination of the most influential features responsible for the bias. This method can be used in a wider range of models, including generalised linear models, and can reduce and refine the amount of information to be eliminated from the embeddings.

The LACE method consists of a minimax game in which the minimax algorithm iteratively finds weights for a classifier that minimises the maximum loss over all possible adversary erasure patterns. Below, we will first construct the loss function to be minimised, and then identify the structure of the minimax game.

Using the same definitions as before, the loss function used in the LACE algorithm is a combination between W_a^c and the predictor of W_a^c, namely $\hat{W_a^c}$. In accordance with Ravfogel et al. (2022), the predictor in our method is constructed using a Generalised Linear Model (GLM) with a logistic link function, such that

$$\hat{W_a^c} = g^{-1}(\theta^{c\intercal} W_t^c) = \frac{\exp(\theta^{c\intercal} W_t^c)}{1 + \exp(\theta^{c\intercal} W_t^c)}, \tag{8}$$

where θ^c contains the model's parameters.

Then, if we consider the log loss function, we can define the loss function for the GLM which predicts W_a^c from W_t^c as follows,

$$\lambda(W_a^c, \hat{W}_a^c) = \lambda(W_a^c, g^{-1}(\theta^{c\intercal} W_t^c)) = W_a^c \log \frac{\exp(\theta^{c\intercal} W_t^c)}{1 + \exp(\theta^{c\intercal} W_t^c)}. \tag{9}$$

Minimising $\lambda(W_a^c, \hat{W}_a^c)$ with respect to θ^c provides a predictor of W_a^c in terms of W_t^c. Now, using the same ideas as in Sec. 4.2, we define P^c to be the orthogonal projection matrix which projects onto the orthogonal complement of the bias subspace of attribute c, of the embeddings. Furthermore, we specify the set Π to contain all projection matrices which neutralise a subspace of the embedding vectors, such that $P^c \in \Pi$. Now we can give a definition for the minimax game which, if solved, gives us the optimal projection matrix for bias reduction of sensitive attribute c as follows,

$$\min_{\theta^c \in \mathbb{R}^{\mathbb{D}}} \max_{P^c \in \Pi} W_a^c \log \frac{\exp(\theta^{c\intercal} P^c W_t^c)}{1 + \exp(\theta^{c\intercal} P^c W_t^c)}. \tag{10}$$

Optimising this minimax game, therefore, requires the minimisation of the objective function with respect to θ^c and the maximisation with respect to P^c simultaneously. By minimising with respect to θ^c, the algorithm searches for a set of coefficients which best explain W_a^c from W_t^c. This leads to a reduced dimension representation of the influence between the two sets of words. On the other hand, maximising with respect to P^c requires the algorithm to select the set of most informative features that can best preserve the information in this same influence between the sets of words. Therefore, the algorithm ultimately searches for the optimal coefficient matrix, θ^c, which explains W_a^c in terms of a reduced set of features, while simultaneously selecting the most informative features to choose from. Simultaneously, the algorithm finds a projection matrix which identifies and removes the most influential features responsible for the influence of W_t^c on W_a^c. In other words, the algorithm finds a projection matrix that minimises the worst-case prediction error, subject to a constraint on the search space for the projection matrix. The outcome for P^c, thus, identifies and removes the most influential features responsible for the bias in W_t^c, while minimising the impact on the prediction $\hat{W}_a^c$. This supplies us with a good choice for the projection matrix, which removes the bias related to c from the embeddings, without removing too much of the influence between the sets of words.

Solving a minimax game, however, is not always easy, as the game is based around finding the best strategy for one player, while taking into account a worst-case scenario for the other player. The optimisation of the overall objective function, therefore, can be very complex and slow. A case in which the optimisation of such objective function is generally well behaved, though, is when the outer optimisation problem is concave, and the inner is convex (Tuy, 2004). A concave-convex minimax game allows for the use of powerful optimisation techniques, such as Alternating Gradient Descent (AGD), for solving the game. Ravfogel et al. (2022) show that the set Π, which defines the search space for the projection matrix is non-convex. Therefore, the researchers propose the Relaxed-LACE (R-LACE) algorithm, which provides a relaxation on the search space for the projection matrix, making it convex. Essentially, this relaxation involves the alteration from the

search space being Π to the search space being the convex hull, conv(Π), of Π. Finally, the concave-convex minimax game to solve is

$$\min_{\theta^c \in \mathbb{R}^{\mathbb{D}}} \max_{P^c \in \text{conv}(\Pi)} W_a^c \log \frac{\exp(\theta^{c\intercal} P^c W_t^c)}{1 + \exp(\theta^{c\intercal} P^c W_t^c)}. \tag{11}$$

Following Ravfogel et al. (2022), our minimax algorithm for solving the game consists of Alternating Stochastic Gradient Descent (ASGD) steps, where we perform alternate minimisation over θ^c and maximisation over P^c. The game, ultimately, yields a projection matrix that is optimised for the attenuation of bias direction c from W_t.

4.4 Multi-Sensitive Attribute Implementations of the Methods in NLP Models

The methods described in Sec. 4.1, Sec. 4.2 and Sec. 4.3 provide us with several techniques for bias attenuation of word embeddings. In this section, the manner in which all methods are included into the contextualised NLP models is described.

As briefly mentioned in Sec. 4.1, the loss function debiasing method is incorporated into the NLP models by adding the constructed loss function to the training objective of the model. This way, the training objective of the model is extended to include a penalty for stereotypical associations. The loss terms for the various sensitive attributes are summed in the penalty term in order to produce an MSA debiasing loss. The advantage of including the debiasing method in the training objective of the models is that we can manipulate all hidden layers of the NLP model and, thus, 'teach' the embeddings to exclude relationships which society would view as discriminatory within its normal training procedure.

With regard to the INLP and R-LACE debiasing methods, explained in Sec. 4.2 and Sec. 4.3, we construct projection matrices which can be applied to NLP models in a 'post-processing' sense. The implementation of these projection matrices is done by multiplying the last hidden state of the word embedding obtained using the NLP model by the projection matrices corresponding to each of the sensitive attributes. This way, the stereotypical associations are removed from the hidden states of the contextualised word embeddings. Each projection matrix corresponds to the bias direction of one sensitive attribute. Multiplying the hidden states with each of the matrices consecutively, projects the embeddings onto the orthogonal complement of the bias subspace, which consists of the several bias directions. These hidden states then pass through the last linear layer of the specified NLP model to result in the eventual predictions of the model in its specific task. An advantage of debiasing using projection matrices is that the matrices are trained separately from the NLP model. This way, obtaining an optimal projection matrix, which minimises stereotypical associations, is the main task of the debiasing method.

Ultimately, in this research, we investigate the performance of the three MSA debiasing methods. We primarily analyse the ability of the debiasing methods to reduce stereotypical associations in NLP models. In addition, we look at the performance of the methods in terms of the accuracy vs. fairness trade-off. Furthermore, the influence of the different debiasing methods for three contextual NLP models is investigated. The NLP models used are BERT, DistilBERT, and ALBERT.

4.5 Performance Measures

For the evaluation of the debiasing methods in each of the models, we use research by Nadeem et al. (2021). The researchers construct performance measures which evaluate MSA stereotypical biases in language models as well as the language modelling ability of the models. Using their research, we can give an indication of how well the debiasing methods, described in the sections above, perform in attenuating stereotypical associations, and in retaining semantic information.

The evaluation measures all use the 'probability' scores which the NLP models output in the Masked Language Modelling (MLM) task of the model. Given the task of MLM, NLP models assign a probability score to each of the words in its vocabulary. This score identifies the probability that the masked word in the input sentence corresponds to the word in the vocabulary. In a general application of MLM, the word in the vocabulary which receives the highest probability score is outputted to correspond to the masked word in the input. In our performance evaluation, however, we focus on the specific score given to the various words in the model's vocabulary itself.

As described in Sec. 3.2, the evaluation data consists of sentence 'triplets'. For the performance evaluation, the context, or association which is different across the different sentences is inputted into the model as the masked token. Then, the probability scores of the three different contexts/associations are saved and used to compute the evaluation metrics. Using the example from Sec. 3.2, the input sentence into the NLP model is "Our housekeeper is a [MASK]", and the scores outputted by the model for the words "Mexican", "American", and "fish" are registered.

Using the scores for the different associations for each of the sentence triplets in the evaluation dataset, the evaluation metrics can be calculated. Ultimately, in correspondence with Nadeem et al. (2021), the evaluation metrics used are the Language Modeling Score (LMS), the Stereotype Score (STS), and the Idealised Context Association Test Score (ICAT). In addition, we propose a Stereotype Proximity Score (SPS), which is used as a normalised alternative for the STS.

The LMS describes the model's ability to correctly identify target words, and is, thus, a representation of the model's accuracy. Using the model output scores for the different instantiations of the sentences, the LMS is calculated as the percentage of times that the NLP model ranks the meaningful associations to be more preferable than the meaningless one across all sentence triplets in the dataset. The LMS is defined as follows,

$$LMS = \frac{\sum_{k=1}^{K} \Big(\mathrm{I}(P_k^{stereo} > P_k^{meaningless}) + \mathrm{I}(P_k^{anti-stereo} > P_k^{meaningless}) \Big)}{2K} \times 100\%, \tag{12}$$

where P_k^{stereo}, $P_k^{anti-stereo}$, and $P_k^{meaningless}$ refer to the probability scores of the stereotypical, anti-stereotypical and meaningless associations, respectively. Furthermore, $k = 1, ..., K$ is the index of the sentence triplet at hand, where K is the total number of sentence triplets. $\mathrm{I}(\cdot)$ refers to the indicator function, such that the sum over all sentence triplets provides a count of the number of times that the stereotypical and anti-stereotypical associations are ranked higher than the meaningless one. The LMS, although not being

a perfect measure for accuracy, gives good insight into the difference in ability to recognise and analyse relationships in language. It, thus, provides us with a metric to identify differences in accuracy before and after debiasing an NLP model.

The STS shows the inclination of the model to prefer stereotypical associations over anti-stereotypical associations. It is calculated as the percentage of times that the NLP model ranks the stereotypical association higher than the anti-stereotypical association, using the scores as described above. The STS is formally defined as

$$STS = \frac{\sum_{k=1}^{K} \mathrm{I}(P_k^{stereo} > P_k^{anti-stereo})}{K} \times 100\%. \tag{13}$$

Note that, whereas an ideal NLP model would have an LMS of 100, the STS of an ideal NLP model is 50. An STS of 50, namely, indicates that the model is indifferent between stereotypical and anti-stereotypical associations for target terms, which means it has no discriminatory inclination towards either side. The SPS, then, is defined as the proximity to a perfectly unbiased model and is a normalisation of the STS. The SPS is calculated as the percentage of absolute difference from the ideal score of 50,

$$SPS = \left(1 - \frac{|STS - 50|}{50}\right) \times 100\%. \tag{14}$$

Lastly, the ICAT is a combination of the LMS and the SPS and is, thus, used as an 'overview' of how well the model performs, both in terms of accuracy and in terms of bias. If we, following Nadeem et al. (2021), assume equal importance between accuracy and fairness, the ICAT is calculated as follows,

$$ICAT = \frac{LMS + SPS}{2}. \tag{15}$$

Although this ICAT (as well as the LMS and SPS) is not a perfect measure for the overall performance of the NLP model, the ICAT score gives us a basis for comparison of the various models. Therefore, it does give us an indication of the performance of the various debiasing methods, as well as their performance in comparison with a non-debiased NLP model.

5 Results

In this section, the results of the debiasing methods are presented and discussed. The performance measures are presented for the various debiased BERT, DistilBERT, and ALBERT models, using each of the debiasing methods described in Sec. 4. In addition, the same performance measures are given for the original, un-debiased BERT, DistilBERT and ALBERT models to serve as baseline results.

In Sec. 5.1, the overall performance of the various non-debiased and debiased NLP models is given. This section gives a comprehensive overview of the differences in accuracy and fairness levels of the different versions of the models. Sec. 5.2, then, presents the performance of the models with regards to the specific sensitive attributes which were debiased. This section shows the models' abilities to attenuate bias of multiple sensitive attributes simultaneously. Finally, Sec. 5.3 gives an analysis of the performance of the various methods with regards to the inevitable accuracy vs. fairness trade-off.

5.1 Overall Model Performance in Relation to MSA Debiasing Methods

Table 1 shows the overall performance results of all versions of the NLP models which were researched. The 'original' versions of the models correspond to the pre-trained, un-debiased versions of the various contextualised NLP models. The term 'Loss' refers to the weighted loss function debiasing method as described in Sec. 4.1. The terms 'INLP' and 'R-LACE' refer to the projection matrix methods described in Sec. 4.2 and Sec. 4.3, respectively. For clarity and comparison, the theoretical performance measure scores for an ideal Language Model (LM), a maximally biased (stereotyped) LM and a random LM are given in the table as well. Ultimately, the closer the performance scores of the debiased NLP models are to those of the theoretically ideal LM, the better the overall performance of the method is.

Table 1. Overall performance of the various language models

Model	**LMS Score**	**SPS Score**	**ICAT Score**
Ideal LM	100	100	100
Stereotyped LM	-	0.0	0.0
Random LM	50.0	100	75.0
BERT-original	**84.7**	80.6	82.7
BERT-debiased (Loss)	76.1	88.6	**82.4**
BERT-debiased (INLP)	59.0	96.8	77.9
BERT-debiased (RLACE)	59.8	**97.8**	78.8
DistilBERT-original	**85.5**	78.2	81.9
DistilBERT-debiased (Loss)	72.8	90.6	**81.7**
DistilBERT-debiased (INLP)	55.2	96.6	75.9
DistilBERT-debiased (RLACE)	54.5	**97.6**	76.1
ALBERT-original	**89.6**	82.0	85.8
ALBERT-debiased (Loss)	68.5	97.2	**82.6**
ALBERT-debiased (INLP)	50.8	98.0	74.4
ALBERT-debiased (RLACE)	53.2	**98.8**	76.0

In Table 1, the highest LMS and SPS scores for each of the NLP models are in bold. For the ICAT scores, the bold scores indicate the optimal value across the debiased versions of the NLP model. Immediately, we see the same trend for the performance scores across the various NLP models. For each of the models, the debiasing technique which leads to the highest ICAT score is the Loss method. The method which reduces the bias in the model the most, across all models, is the R-LACE projection matrix method. Furthermore, all debiased NLP models exhibit equal scale increases in SPS compared to the original version, across the various language models. These results give an initial indication of universality for the debiasing methods. Across the different NLP models, each of the debiasing methods manages to decrease the bias in the model.

Another result we notice is the inevitable decrease in language modelling ability of the debiased models with respect to the original models. For each of the NLP models, the LMS of the original model is clearly higher than the debiased versions of the model. This, however, is unavoidable and at the basis of the 'accuracy vs. fairness' trade-off. By adjusting NLP model training structures, or post-processing the output using projections, we are slightly shifting the model's objective. Instead of aiming for pure accuracy of the NLP model output, we force the model to take into account bias structures in the data. This causes the model to lose some of its accuracy, in return for some increase of fairness in the model.

In general, the results show that, whilst the projection matrix methods decrease bias (increase the SPS) in the models the most, they also cause a greater decrease in accuracy than the Loss method. This is to be expected, as more of the model's 'focus' switches from accuracy towards debiasing stereotypical associations. The fact that the INLP and R-LACE methods are more extreme in reducing bias, and thus in decreasing accuracy of the model, can be explained by the structure of the debiasing method. These methods construct projection matrices independently from the training objectives of the original model. Therefore, these debiasing techniques barely contain regularisations with respect to semantic abilities of the specific NLP model. The Loss method, on the other hand, explicitly includes a regularisation term which focuses on retaining a certain level of accuracy of the original model.

Zooming in on the differences between the two projection matrix methods, we see that the R-LACE method captures slightly more of the bias than the INLP method, in each of the models. In addition, we notice that in the BERT model, the R-LACE method performs better than the INLP method, with regards to both the LMS and the SPS. This superiority might be explained by the R-LACE method's more diverse nature. This method allows for the stereotypical influences to be captured using Generalised Linear Models, causing there to be more flexibility in identifying and attenuating the bias subspace than allowed for by the INLP method. In addition, we note that the focus of the R-LACE method is more explicitly and specifically on the minimisation of information loss than the focus of the INLP method is.

Finally, we can label the Loss method to be closer to a compromise of accuracy and fairness than the projection methods. This debiasing method is added to the NLP models as an extra penalty in its training objective, allowing most of the focus of the model to remain on the semantic information retention. Although the SPS of the Loss method, in each of the models, is lower than the SPS of the other debiasing methods, we can still see a clear trend in the reduction of bias with respect to the original model. In addition, the LMS scores for this debiasing technique are distinctly higher than the LMS scores of the other two debiasing methods.

Based on these overall results, we might conclude that, out of the debiasing techniques, the Loss method performs the best with regards to the accuracy vs. fairness trade-off. The same can be concluded from the ICAT scores of each of the models. Assuming equal importance between accuracy and fairness, the Loss method scores distinctively higher in overall performance score than both the projection methods. With regards to pure bias reduction in NLP models, however, the INLP and R-LACE methods perform better

than the Loss method.

5.2 Sensitive Attribute-Specific Model Performance for the MSA Methods

The focus of this paper, in addition to overall bias attenuation in contextualised NLP models, is on the mitigation of bias related to several sensitive attributes simultaneously. Therefore, in Table 2, the performance scores given in Table 1 are divided into the performance within the specific sensitive attributes, which are attenuated by the methods. Using these results, we can analyse the MSA debiasing methods' abilities to reduce bias for specific sensitive classes simultaneously, and perform some sensitivity analysis of the methods with respect to the various attributes. As described in Sec. 3.2, the data used for performance evaluation can be split up into four different demographics and, therefore, we can evaluate the performance scores for each of these demographics independently. These demographics directly relate to attributes which are prone to sensitive biases. The demographics are gender, occupation, race, and religion. Table 2 contains the LMS, SPS, and ICAT scores for each of the instantiations of the NLP models across the various sensitive attribute classes.

Table 2. Per attribute performance of the various MSA debiased language models

		Gender			Occupation			Race			Religion		
		LMS	SPS	ICAT	LMS	SPS	ICAT	LMS	SPS	ICAT	LMS	SPS	ICAT
BERT	Orig	**86.0**	72.2	79.1	**82.7**	77.2	80.0	**85.7**	85.2	85.5	**88.5**	87.0	87.8
	Loss	84.5	83.4	**84.0**	77.2	92.0	**84.6**	73.1	86.4	**79.8**	74.4	99.2	**86.8**
	INLP	60.5	86.4	73.5	60.1	93.6	76.9	57.8	**98.4**	78.1	57.5	91.6	74.6
	RLACE	61.8	**89.0**	75.4	60.7	**94.6**	77.7	58.7	97.4	78.1	58.0	**96.6**	77.3
DBERT	Orig	**86.4**	79.8	83.1	**83.4**	71.8	77.6	**86.8**	83.4	85.1	**89.1**	76.0	82.6
	Loss	83.2	81.0	**82.1**	72.1	93.2	**82.7**	70.5	90.8	**80.7**	74.4	**92.2**	**83.3**
	INLP	58.4	91.2	74.8	55.7	95.2	75.5	53.3	**99.2**	76.3	60.9	78.2	69.6
	RLACE	59.5	**92.4**	76.0	54.5	**97.2**	75.9	52.7	98.8	75.8	59.6	76.3	68.0
ALBERT	Orig	**89.3**	80.0	84.7	**88.9**	79.2	84.1	**90.2**	83.0	86.6	**92.7**	79.4	86.1
	Loss	85.2	90.0	**87.6**	69.1	94.6	**81.9**	62.7	**98.6**	**80.7**	81.3	**92.8**	**87.1**
	INLP	53.8	95.8	74.8	52.7	98.6	75.7	48.8	96.6	72.7	45.4	89.6	67.5
	RLACE	55.4	**98.8**	77.1	55.0	**98.8**	76.9	51.6	93.4	72.5	46.5	90.0	68.3

Generally, we can see similar trends to the ones in Table 1, for each of the attribute categories. Specifically, we should note that there is an increase in SPS for all categories in each of the models for all debiasing methods with respect to the original models. This indicates that each of the debiasing methods is able to attenuate bias in multiple sensitive attributes simultaneously. However, we do observe some variation in increase of SPS across the categories, which indicates that the methods have varying influence on the biases. The religion attribute, for example, has only small increase in SPS for the projection methods in the DistilBERT model. This could be due to the fact there there was too little training data for this sensitive attribute to construct an accurate projection matrix, or that this type of bias is difficult to capture using a projection matrix.

All in all, Table 2 gives us evidence that the debiasing techniques are able to mitigate bias in an MSA framework. Each of the sensitive attributes shows an increase in SPS for the debiased models, compared to the original models. Some attributes show greater increase than others, but the overall trend shows promising results. The variations in LMS and SPS across categories can be due to difference in training data availability for that category, NLP model architectures or even the structure of the stereotypical associations in that category.

5.3 Accuracy vs. Fairness Trade-Off

Overall, we can conclude that each of the bias mitigation techniques described in Sec. 4 is able to attenuate MSA bias in contextualised word embeddings. The SPS scores in the tables in Sec. 5.1 and Sec. 5.2 all show increased values for the various debiased models compared to the original models. This indicates that each of the debiasing methods reduces the corresponding model's tendency to prefer stereotypical associations over anti-stereotypical associations in multiple sensitive attribute classes. The different debiasing techniques, however, do show differences in the values for the SPS, as well as for the LMS. The INLP and R-LACE projection methods increase the SPS more than the Loss method does. However, these models also cause lower LMS scores than the Loss method.

These results demonstrate some of the limitations of each of the models. The INLP and R-LACE methods are post-processing methods, meaning that they modify the model embeddings by projecting them onto a new subspace. This process can lead to semantic information loss of the NLP model, especially for low-frequency words. In comparison, the Loss method modifies the model embeddings in a way which preserves semantic information. This method, however, also has its limitations. Residual bias, for example, can exist in the final results, as the focus on semantic information retention is distinctively present.

If we were to ignore the accuracy vs. fairness trade-off, we could conclude that the INLP and R-LACE debiasing techniques are the superior techniques for bias attenuation in contextualised NLP models. However, the retention of semantic information of the NLP model is important in practice as well, and the results in the tables also show that these methods lose a fair amount of the language modelling abilities of the models. In some categories (e.g. the race attribute in the DistilBERT model) the LMS of the INLP and RLACE debiased models are reduced almost to the point where the LMS resembles that of a random LM. Therefore, in most applications, the Loss method would be preferred. In each attribute instance, and in each NLP model type, the LMS remains distinctively above the level of that of a random LM. This research indicates, therefore, that the Loss debiasing method is able to mitigate MSA biases in contextualised NLP models whilst still retaining a level of semantic information of the original model.

6 Conclusion

With the increase in everyday use of NLP-based applications, the importance of recognising and attenuating problematic biases embedded in these applications grows. NLP has grown to become an integral aspect of our daily tasks, but NLP models, through attempts of capturing human semantics, inevitably encode human social biases as well. Recent research (e.g. Bolukbasi et al., 2016) has shown that such bias can lead to discriminatory practices which are unacceptable from our current social, ethical, and even legal perspectives. The need for research into bias mitigation, therefore, becomes bigger and bigger.

Whilst most research into this topic has focused on the attenuation of bias from static word embeddings, contextualised word embeddings have been established as the state-of-the-art 'type' of embedding for NLP tasks. Research has shown, unfortunately, that such contextualised word embeddings encode unfair bias as well. In addition, recent research has focused on the mitigation of bias from NLP models with respect to a single sensitive stereotyped attribute. In contrast, real data generally inherits bias with regards to multiple sensitive attributes, rather than merely one.

This paper, therefore concentrates on providing new insights into possibilities of the mitigation of MSA bias from contextualised textual frameworks, with a focus on the main research question: *How can we attenuate bias in textual analysis whilst retaining an appropriate level of semantic information?*

The paper presents three methods which can provide novel, pragmatic approaches for the MSA attenuation of stereotypical associations in state-of-the-art NLP models and applications. The first method (the 'Loss' method) is based around extending the loss function in the NLP model structure by adding a penalty term to the training objective. This alternative training objective, then, penalises stereotypical associations, whilst simultaneously regulating an appropriate level of retention of semantic information in the model. The second and third methods (the INLP and R-LACE methods) require the construction of a projection matrix, which projects the NLP model output onto the orthogonal complement of the bias subspace. The difference between these two methods is in the construction of the projection matrix. The INLP technique structures the projection matrix by using an iterative algorithm which updates the projection matrix so that more and more of the stereotypical influence is captured after each step. The R-LACE method revolves around solving a minimax game which, if solved, finds the projection matrix that minimises the worst-case prediction error between sensitive attributes and stereotypical outcomes.

Ultimately, the results show that all methods are capable of reducing MSA bias to a distinctive extent. The projection matrix methods reduce the bias in the models more than the Loss method does. However, these methods also decrease the language modelling capabilities of the NLP models greatly. Although both the INLP and R-LACE methods show the ability to reduce the bias almost to the level of that of an ideal language model, they, in some attribute categories, reduce the language modelling ability towards the level of a random language model. Taking into account a decent trade-off between accuracy and fairness of the NLP models, therefore, leads to the Loss method providing the most promising results. This method attenuates bias of multiple sensitive attributes simultaneously, whilst retaining a level of semantic information inherent

in the original, un-debiased model. The results show that bias is reduced in all attribute categories in all NLP models tested, indicating a clear capability of bias mitigation. In addition, the level of retention of language modelling ability of this method is much higher than that of the projection matrix methods.

For future research, it could prove beneficial to tune the hyperparameters of the NLP models, as well as of the models used in the debiasing methods. Better tuned models and debiasing techniques could provide better insight in the full potential of the different methods. In addition, it might be interesting to investigate manners in which a better regularisation component can be added to the INLP and R-LACE methods in order to allow for better information retention.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.

Biddle, S. (2022). The internet's new favorite ai proposes torturing iranians and surveilling mosques.

Black, J. S. and van Esch, P. (2020). Ai-enabled recruiting: What is it and how should a manager use it? *Business Horizons*, 63(2):215–226.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Bommasani, R., Davis, K., and Cardie, C. (2020). Interpreting pretrained contextualized representations via reductions to static embeddings. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Bowman, S. R., Potts, C., and Manning, C. D. (2015). Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 12–21, Beijing, China. Association for Computational Linguistics.

Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., and Banaji, M. R. (2022). Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*.

Dev, S., Li, T., Phillips, J. M., and Srikumar, V. (2020). On measuring and mitigating biased inferences of word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7659–7666.

Dev, S., Li, T., Phillips, J. M., and Srikumar, V. (2021). Oscar: Orthogonal subspace correction and rectification of biases in word embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Dev, S. and Phillips, J. M. (2019). Attenuating bias in word vectors. *CoRR*, abs/1901.07656.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56(1):5–18.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Dhami, D. (2020). Understanding bert-word embeddings.

EMNLP (2021). News-commentary-v15.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16).

Guynn, J. (2015). Google photos labeled black people 'gorillas'. *USA Today Tech*.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Kamiran, F., Karim, A., Verwer, S., and Goudriaan, H. (2012). Classifying socially sensitive data without discrimination: An analysis of a crime suspect dataset. pages 370–377.

Kaneko, M. and Bollegala, D. (2019). Gender-preserving debiasing for pre-trained word embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Kaneko, M. and Bollegala, D. (2021). Debiasing pre-trained contextualised embeddings. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Lahoti, P., Gummadi, K. P., and Weikum, G. (2019). Ifair: Learning individually fair data representations for algorithmic decision making. *2019 IEEE 35th International Conference on Data Engineering (ICDE)*.

Lohia, P. K., Natesan Ramamurthy, K., Bhide, M., Saha, D., Varshney, K. R., and Puri, R. (2019). Bias mitigation post-processing for individual and group fairness. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Manzini, T., Yao Chong, L., Black, A. W., and Tsvetkov, Y. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *Proceedings of the 2019 Conference of the North*.

Nadeem, M., Bethke, A., and Reddy, S. (2021). Stereoset: Measuring stereotypical bias in pretrained language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Nalisnick, E., Mitra, B., Craswell, N., and Caruana, R. (2016). Improving document ranking with dual word embeddings. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*.

Nielsen, L. (2021). 10 use-cases in everyday business operations using nlp.

OpenAI (2021). Chatgpt. `https://github.com/openai/gpt-3/blob/main/models/chatgpt/README.md`. Accessed on March 24th, 2023.

Popović, R., Lemmerich, F., and Strohmaier, M. (2020). Joint multiclass debiasing of word embeddings. *Lecture Notes in Computer Science*, page 79–89.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Ravfogel, S., Twiton, M., Goldberg, Y., and Cotterell, R. D. (2022). Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR.

Schmidt, B. (2015). Rejecting the gender binary: a vector-space operation.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Tuy, H. (2004). Minimax theorems revisited. *Acta Mathematica Vietnamica*, 29.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. (2017). Fa*ir: A fair top-k ranking algorithm. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W. (2019). Gender bias in contextualized word embeddings. *Proceedings of the 2019 Conference of the North*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018a). Gender bias in coreference resolution: Evaluation and debiasing methods. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K.-W. (2018b). Learning gender-neutral word embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Zuo, Y., Gong, C., Zhang, W., Yu, P. S., and Liu, W. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1951–1954. ACM.

A Overview of Definitions

Table 3. Overview of the variables used in Sec. 4, with corresponding definitions

Variable	Definition
w	Word or token
c	Sensitive attribute class (e.g. gender, race)
Λ_{a_c}	Set of attribute words of sensitive class c (where $c = 1, ..., C$) (e.g. "he", "she", "Muslim", "Asian")
Λ_{t_c}	Set of target words (e.g. occupations, polarised adjectives); these words are expected to be neutral with respect to the attribute words
W_{a_c}	The set of all embeddings of $w \in \Lambda_{a_c}$
W_{t_c}	The set of all embeddings of $w \in \Lambda_{t_c}$
$\Omega(w)$	Set of sentences containing word (or token) w
$A_c = \bigcup_{w \in \Lambda_{a_c}} \Omega(w)$	Set of sentences containing one or more attribute words
$T_c = \bigcup_{w \in \Lambda_{t_c}} \Omega(w)$	Set of sentences containing one or more target words
E	Embedding model
θ_e	Pre-trained parameters of embedding model e
x	Input sentence
$E_i(w, x; \theta_e)$	Embedding of token, w, in sentence, x, in the i-th layer of embedding model E (where $i = 1, ..., N$)
$\nu_i(w)$	Non-contextualised embedding of w in the i-th layer; taken as the average of the contextualised embeddings of w in the i-th layer of E over all Ω_w
M^c	The bias subspace of sensitive attribute c
$N(M^c)$	The null-space (orthogonal complement of the row-space) of M^c
B^c	The basis vectors of $N(M^c)$
P^c	Projection matrix which projects W_t^c onto the null-space of its bias subspace
Z^c	Alternative representations of W_t^c
$\hat{W}_a^c$	The predictor of W_a^c
θ^c	GLM model parameters
$\lambda(W_a^c, \hat{W}_a^c)$	Log loss function of W_a^c and its predictor
Π	The set of all projection matrices which neutralise a subspace of the embeddings
conv(Π)	The convex hull of Π

B Summary of Attribute and Target Word Lists

B.1 Sets of Attribute Words

Gender Attribute Words:

man, gentleman, guy, woman, lady, girl, him, her, his, hers, he, she, businessman, businesswoman, priest, nun, mr., mrs., daughter, son, father, mother, councilman, councilwoman, etc.

Ethnicities Attribute Words:

aborigine, african, african-american, asian, black, caucasian, hispanic, white, brown, dark, light, etc.

Religion Attribute Words:

anglican, atheist, atheists, buddhist, catholic, christian, hindu, jew, jewish, morman, muslim, protestant, sikh, christianity, judaism, islam, buddhism, etc.

B.2 Sets of Target Words

Gender Stereotype Target Words:

Occupations: trucker, soldier, astronomer, skipper, banker, dancer, educator, chef, hairdresser, homemaker, police, footballer, nanny, nurse, etc.

Character Traits: jealousy, lovely, geek, gentle, nurturing, leader, soft, sweet, hysterical, aggressive, etc.

Others: warrior, beauty, flirt, flower, make-up, jewelry, mirror, etc.

Polarised and Class Target Words:

Polarised: aggressive, arrogant, clean, clever, cruel, evil, dumb, honest, intelligent, rude, professional, smart, strong, ugly, weak, murderer, champion, terrorist, thief, etc.

Class: affluent, impoverished, needy, poor, prosperous, rich, underpriviliged, etc.

Physical Target Words:

athletic, fat, huge, lanky, little, muscular, short, slender, slim, tall, voluptuous, etc.