

Exploration and Mitigation of Gender Bias in Word Embeddings from Transformer-based Language Models

by

Ariyan Hossain
20101099

Rakinul Haque
20101290

Khondokar Mohammad Ahanaf Hannan
20101079

Nowreen Tarannum Rafa
20101329

Humayra Musarrat
20101089

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
School of Data and Sciences
September 2023

© 2023. Brac University
All rights reserved.

Declaration

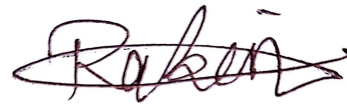
It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Ariyan Hossain
20101099



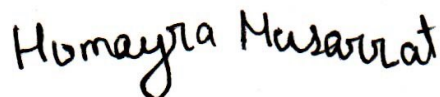
Rakinul Haque
20101290



Khondokar Mohammad Ahanaf Hannan
20101079



Nowreen Tarannum Raza
20101329



Humayra Musarrat
20101089

Approval

The thesis/project titled “Exploration and Mitigation of Gender Bias in Word Embeddings from Transformer-based Language Models” submitted by

1. Ariyan Hossain (20101099)
2. Rakinul Haque (20101290)
3. Khondokar Mohammad Ahanaf Hannan (20101079)
4. Nowreen Tarannum Rafa (20101329)
5. Humayra Musarrat (20101089)

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on September, 2023.

Examining Committee:

Supervisor: (Member)



Dr. Farig Yousuf Sadeque
Assistant Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor: (Member)



Shayekh Bin Islam
Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor: (Member)



Shoaib Ahmed Dipu
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator: (Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department: (Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Machine learning has the potential to uncover data biases resulting from human error when it's implemented without proper restraint. However, this complexity arises from word embedding, which is a prominent technique for capturing textual input as vectors applied in different machine learning and natural language processing tasks. Word embeddings are biased because they are trained on text data, which frequently incorporates prejudice and bias from society. These biases may become deeply established in the embeddings, producing unfair or biased results in AI applications. There are efforts made to recognise and lessen certain prejudices, but comprehensive bias elimination is still a difficult task. In Natural Language Processing (NLP) systems, contextualized word embeddings have taken the place of traditional embeddings as the preferred source of representational knowledge. It is critical to evaluate biases contained in their replacements as well since biases of various kinds have already been discovered in standard word embeddings. Our focus is on transformer-based language models, primarily BERT, which produce contextual word embeddings. To measure the extent to which gender biases exist, we apply various methods like cosine similarity test, direct bias test and ultimately detect bias through probability of filling MASK by the models. Based on this probability, we develop a novel metric called MALoR to observe bias. Finally, to mitigate the bias, we continue pretraining these models on a gender balanced dataset. Gender balanced dataset is created by applying Counterfactual Data Augmentation (CDA). To ensure consistency, we perform our experiments on different gender pronouns and nouns - "he-she", "his-her" and "male names-female names". These debiased models can then be used across several applications.

Keywords: Natural Language Processing; Gender Bias; Debiasing; Word embeddings; BERT; Continued Pretraining

Acknowledgement

To begin, we sincerely express our thankfulness and praise to Allah, whose divine direction has enabled us to conclude our thesis without severe setbacks. Second, we would like to thank our supervisor, Dr. Farig Yousuf Sadeque, as well as our co-supervisors, Shayekh Bin Islam and Shoaib Ahmed Dipu, for their continuous support and essential guidance during our research journey. They have had an important role in developing our work. Finally, we would like to express our heartfelt gratitude to our parents for their constant support and prayers, which have been the foundation of our successes, bringing us to the edge of graduation. The completion of our thesis has been a journey filled with heavenly blessings, mentoring, and constant family support. We recognise and regard each of these factors as essential to our success, and we look forward to using the lessons and principles we have learned from this experience in our future endeavours with great humility and appreciation.

Table of Contents

Declaration	i
Approval	ii
Abstract	iv
Acknowledgment	v
Table of Contents	vi
List of Tables	vii
List of Figures	1
1 Introduction	2
1.1 Introduction	2
1.2 Problem Statement	3
1.3 Research Objectives	4
2 Related Work	5
2.1 Static Word Embedding	5
2.2 Contextualized Word Embedding	12
3 Dataset	18
3.1 Work Flow	18
3.2 Winogender and Direct Bias	19
3.3 Sentence Structures	19
3.3.1 Structures for gendered term - he and she	20
3.3.2 Structures for gendered term - his and her	22
3.3.3 Structures for gendered term - male names and female names	24
3.4 Data Extraction and Augmentation	26
3.4.1 Counterfactual Data Augmentation (CDA)	26
3.4.2 Dataset for gendered pronoun - he and she	27
3.4.3 Dataset for gendered pronoun - his and her	27
3.4.4 Dataset for gendered pronoun - male names and female names	27
4 Methodology	29
4.1 Model Description	29
4.1.1 Introduction of BERT	29
4.1.2 Variations of BERT	31

4.2	Cosine Similarity Test	32
4.3	Direct Bias Test	33
4.4	Masking Probability of BERT	35
4.4.1	Masking the gendered term	36
4.4.2	Masking the occupation	38
4.5	Visualization of Masking Probability	39
4.5.1	Masking the gendered term for exp 1: he-she	39
4.5.2	Masking the occupation for exp 1: he-she	40
4.5.3	Masking the gendered term for exp 2: his-her	41
4.5.4	Masking the occupation for exp 2: his-her	42
4.5.5	Masking the gendered term for exp 3: male-female name	43
4.5.6	Masking the occupation for exp 3: male-female name	44
5	Debiasing Technique	45
5.1	Bias Evaluation Metric - MALoR	45
5.1.1	Metric for gendered term - he and she	46
5.1.2	Metric for gendered term - his and her	46
5.1.3	Metric for gendered term - male names and female names	46
5.2	Continue Pre training for Debiasing	47
5.2.1	Preprocessing	47
5.2.2	Training	48
6	Results and Discussion	50
6.1	Learning Graph- Epoch vs MALoR	50
6.2	Results of Bias Evaluation Metric MALoR of models before and after debiasing	52
6.3	Visualization of Masking probability before and after debiasing	53
6.3.1	Masking the gendered term for exp 1: he-she	53
6.3.2	Masking the occupation for exp 1: he-she	54
6.3.3	Masking the gendered term for exp 2: his-her	55
6.3.4	Masking the occupation for exp 2: his-her	56
6.3.5	Masking the gendered pronoun for exp 3: male-female name	57
6.3.6	Masking the occupation for exp 3: male-female name	58
6.4	Quality Assurance of debiased BERT	59
7	Conclusion	60
7.1	Findings and Contributions	60
7.2	Limitations and Future Work	61
	Bibliography	61

List of Tables

3.1	Sentences structures for “he-she”	22
3.2	Sentences structures for “his-her”	23
3.3	Sentences structures for “male names-female names”	26
3.4	Mapping of Common Male Names to Corresponding Female Names .	28
4.1	Difference between different BERT variants [41]	31
5.1	MALoR Scores of different models	47
6.1	Convergence Rates for last 5 epochs	51
6.2	MALoR Scores of he-she	52
6.3	MALoR Scores of his-her	52
6.4	MALoR Scores of male-female	52
6.5	Before and After Comparison of SST-2 Accuracy	59

List of Figures

3.1	Work Flow	18
4.1	BERT architecture [10]	30
4.2	Cosine Similarity Test	33
4.3	Direct Bias Test	35
4.4	Missing Occupations in Vocab	37
4.5	Missing Names in Vocab	38
4.6	BERT Base Mask Gender exp 1	39
4.7	BERT Large Mask Gender exp 1	39
4.8	ALBERT Base Mask Gender exp 1	39
4.9	ALBERT Large Mask Gender exp 1	39
4.10	DistilBERT Base Mask Gender exp 1	39
4.11	RoBERTa Base Mask Gender exp 1	39
4.12	RoBERTa Large Mask Gender exp 1	40
4.13	BERT Base Mask Occ exp 1	40
4.14	BERT Large Mask Occ exp 1	40
4.15	DistilBERT Base Mask Occ exp 1	40
4.16	BERT Base Mask Gender exp 2	41
4.17	BERT Large Mask Gender exp 2	41
4.18	ALBERT Base Mask Gender exp 2	41
4.19	ALBERT Large Mask Occ exp 2	41
4.20	DistilBERT Base Mask Gender exp 2	41
4.21	RoBERTa Base Mask Gender exp 2	41
4.22	RoBERTa Large Mask Gender exp 2	42
4.23	BERT Base Mask Occ exp 2	42
4.24	BERT Base Mask Occ exp 2	42
4.25	DistilBERT Base Mask Occ exp 2	42
4.26	BERT Base Mask Gender exp 3	43
4.27	BERT Large Mask Gender exp 3	43
4.28	DistilBERT Base Mask Gender exp 3	43
4.29	BERT Base Mask Occ exp 3	44
4.30	BERT Large Mask Gender exp 3	44
4.31	DistilBERT Base Mask Gender exp 3	44
6.1	BERT Base Graph exp 1	50
6.2	BERT Base Graph exp 2	50
6.3	BERT Base Graph exp 3	50
6.4	BERT Large Graph exp 1	50
6.5	BERT Large Graph exp 2	50

6.6	BERT Large Graph exp 3	50
6.7	DistilBERT Base Graph exp 1	51
6.8	DistilBERT Base Graph exp 2	51
6.9	DistilBERT Base Graph exp 3	51
6.10	Original BERT Base Mask Gender exp 1	53
6.11	Debiased BERT Base Mask Gender exp 1	53
6.12	Original BERT Large Mask Gender exp 1	53
6.13	Debiased BERT Large Mask Gender exp 1	53
6.14	Original DistilBERT Mask Gender exp 1	53
6.15	Debiased DistilBERT Mask Gender exp 1	53
6.16	Original BERT Base Mask Occ exp 1	54
6.17	Debiased BERT Base Mask Occ exp 1	54
6.18	Original BERT Large Mask Occ exp 1	54
6.19	Debiased BERT Large Mask Occ exp 1	54
6.20	Original DistilBERT Mask Occ exp 1	54
6.21	Debiased DistilBERT Mask Occ exp 1	54
6.22	Original BERT Base Mask Gender exp 2	55
6.23	Debiased BERT Base Mask Gender exp 2	55
6.24	Original BERT Large Mask Gender exp 2	55
6.25	Debiased BERT Large Mask Gender exp 2	55
6.26	Original DistilBERT Mask Gender exp 2	55
6.27	Debiased DistilBERT Mask Gender exp 2	55
6.28	Original BERT Base Mask Occ exp 2	56
6.29	Debiased BERT Base Mask Occ exp 2	56
6.30	Original BERT Large Mask Occ exp 2	56
6.31	Debiased BERT Large Mask Occ exp 2	56
6.32	Original DistilBERT Mask Occ exp 2	56
6.33	Debiased DistilBERT Mask Occ exp 2	56
6.34	Original BERT Base Mask Gender exp 3	57
6.35	Debiased BERT Base Mask Gender exp 3	57
6.36	Original BERT Large Mask Gender exp 3	57
6.37	Debiased BERT Large Mask Gender exp 3	57
6.38	Original DistilBERT Mask Gender exp 3	57
6.39	Debiased DistilBERT Mask Gender exp 3	57
6.40	Original BERT Base Mask Occ exp 3	58
6.41	Debiased BERT Base Mask Occ exp 3	58
6.42	Original BERT Large Mask Occ exp 3	58
6.43	Debiased BERT Large Mask Occ exp 3	58
6.44	Original DistilBERT Mask Occ exp 3	58
6.45	Debiased DistilBERT Mask Occ exp 3	58

Chapter 1

Introduction

1.1 Introduction

There has been much discussion and research about the existence of linguistic bias against one gender or another. There are ways to avoid perpetuating bias for instance, the singular “he” can be used interchangeably with the plural “they” to refer to people of either gender [1]. These days, Natural Language Processing (NLP) systems are trained using massive amounts of data from a variety of sources. According to Shah et al. [31], biases are not only conveyed in these massive datasets through word choice, but also through word frequency or word co-occurrence frequency. For instance, a model would learn to strongly link nurses with the female gender if the majority of occurrences of the word “nurse” in a corpus have female referents and hence co-occur with female pronouns and female first names.

Word embeddings have helped computers recognize text better. Textual data is difficult to work with because computers and machine learning models lack human-level language comprehension. Word embeddings are word representations that bridge human and computer perceptions of words. In language analysis, especially NLP, this term describes how words are represented for text analysis. It’s mostly used with real-valued vectors. It understands words well enough to anticipate that two words next to each other in a vector subspace are semantically similar. Word embedding have been demonstrated to capture important relationships between words by using the vector differences between them.

There are two different approaches to representing words in NLP tasks - Static Word Embeddings and Contextualized Word Embeddings. Static word embeddings give a single vector for a word no matter the context. For example, in the sentence “I love to watch movies” and “I wear watch”, it gives the same embedding to both “watch” in the two sentences. On the other hand, contextualized word embeddings give different vectors to words depending on the context of the sentence. Hence, it will give two different embedding to the word “watch” in the two sentences as they differ in their meaning. Techniques such as GloVe, Word2Vec are used for creating static word embeddings while BERT, GPT are used for creating contextualized word embeddings.

Significant biases in the training data are typically shown by word embeddings.

Research suggests that neural word embedding activities may be gender skewed. It often correlates masculine phrases with scientific keywords and feminine phrases with artistic keywords [11]. According to Zhao et al. [8], word embeddings often magnify the inherent biases in the training data and as a result a frequent instance of bias can be found in machine learning algorithms. Gender-biased models have the potential to perpetuate and intensify existing gender prejudices in the actual world. For example, if a model is utilised for predicting which job candidates would be successful and it is biased against women, women will be less likely to be hired for those jobs. This can create a vicious cycle in which women are underrepresented in specific industries, making it harder for females to break into such sectors.

Word embeddings have been the subject of hundreds of academic studies. Conversely, very few of these studies has emphasized the severity of the sexism these embeddings have, which risks introducing biases of various sorts into operational systems. In this study, we try to fill in such shortcomings in current studies by creating a comprehensive yet efficient method for reducing sexism in word embeddings. Though there have been a number of attempts to combat the gender disparity present in word embedding, but so far none of them have been very successful and the research is still ongoing. For example, one method designed to examine word embeddings for sexism is the Word Embedding Association Test (WEAT) [7]. May et al. [29] extended this to create Sentence Encoder Association Test (SEAT) to adapt the WEAT to sentences. This helped detecting bias in models creating contextualized word embeddings. While there is much work done on static word embeddings, there is less research done on the bias in contextualized word embeddings. Contextualized word embeddings are vector representations of a word based on the context hence the same word can be represented in different vectors. Due to not having a fixed vector for a word, it is more complex to analyze and remove bias from them. Most of the methods used to work for static word embeddings often don't work for the contextualized word embeddings.

1.2 Problem Statement

An important societal problem that maintains negative perceptions and disparities between men and women is gender bias. It results from deep cultural norms, old customs, and prejudiced mindsets that have formed nations for generations. Gender bias can have serious repercussions in the context of machine learning algorithms, amplifying pre existing biases and producing unjust results. For instance, a study conducted by Buolamwini and Gebru [12] found that widely used facial analysis software made more mistakes when analyzing women's faces compared to men's. Another recent study conducted by Lee et al. [43] revealed that Amazon, where 60% of the workforce is male and 74% of managerial positions are held by men, employed a recruitment algorithm that prioritized word patterns over qualifications. This algorithm, used for evaluating job applicants in a mostly male engineering department, showed bias against resumes containing terms like "women". This bias contributed to ongoing gender discrimination in the hiring process.

Specifically, this work seeks to address the following two research questions:

- RQ1: How can we develop methods to detect gender bias in contextualized embeddings from different transformer models?
- RQ2: How can we effectively mitigate gender bias from these models?

1.3 Research Objectives

While researching the existing debiasing algorithms for word embeddings, we noticed that the majority of them were focused on static word embeddings. Work focusing on removing implicit bias from contextualized word embeddings was scarce. It is a lot more challenging to work with contextualized word embeddings. Biases in transformer models are also quite a new topic, and hence we focus on exploring that. BERT and other pretrained transformers are used to essentially create a representation of their text input for further fine-tuning, and these representations are often superior to word embeddings. But as the good things get amplified here, so do the bad, and gender bias seems to flare up in these pre-trained models as well. In this work, we focus on the following objectives:

- (i) To study the existing bias in static word embeddings and contextualized word embeddings.
- (ii) To look at strategies for identifying bias in contextualized word embeddings.
- (iii) To employ transformer-based language models such as BERT, ALBERT, RoBERTa and DistilBERT which are pre-trained models, to find the existing bias.
- (iv) To focus on eliminating the bias from the models through continuing pretraining.
- (v) To compare our model before and after debiasing to show our success.

Chapter 2

Related Work

2.1 Static Word Embedding

Bolukbasi et al. [6] first addressed the issue of implicit gender imbalance in word embeddings and provided a fix. The authors described a debiasing approach that would reduce unintentional gender biases in embeddings while maintaining their useful characteristics. Their study presented quantitative evidence that word embeddings contained geometric biases. This research also offered two techniques to address this issue: hard and soft debiasing. To begin, they compared a vector representation to the vectors of two gender-specific words to assess any inherent bias. For example, if the vectors of a man and a woman were not at the same distance from the vector of a nurse, it indicated bias. They calculated how far apart or similar two vectors were using cosine similarity. They learned a gender subspace in the embedding using gender-specific phrases like she, he, father, mother, brother, and sister, and the debiasing algorithm will only apply to and eliminate biases from gender-neutral terms while keeping the meanings of gender-specific terms. Because of the small number of gender-distinct terms, they discovered that it was more practical to view the collection of gender-specific words as ‘S’ and the collection of gender-neutral phrases as the complement (N=W/S). They used these words to train a linear SVM classifier to get all gender-specific words. They examined the corpus for gender-neutral terms like programmer, homemaker, and gardener, using all gender-specific phrases. This way, gendered and non-gendered terms were clearly distinguished. They chose the y-axis for the gender-specific paired words. They mapped all words onto the gender axis by starting and finishing with she-he. This step assisted in identifying the direction of embeddings that capture bias. They proposed two methods to remove bias from embeddings: hard debias (Neutralize and Equalize) and soft debias. All gender references from gender-neutral phrases were removed in hard debiasing. Neutralize eliminates gender-neutral words in the gender subspace. To remove gender, they projected all gender-neutral phrases onto the y axis. Equalize guarantees that any unbiased phrase is equally distant from all other phrases in each equality set. By properly equalizing groups of words outside the subdomain from the gender-specific terms (he-she) as the axis, this happens. Equalize eliminates useful disparities, which is its biggest downside. Next, soft debias minimized disparities between both sets while keeping as much resemblance to the initial embedding as attainable, with the degree of similarity regulated by a parameter. Depending on the hyperparameter, soft debias could neutralize gen-

dered phrases. To determine if the solutions met their specific requirements, the authors ran several tests. Analogy generation was used to evaluate their debiasing approach. They polled a crowd after programming a system to generate “she” and “he” phrases. 19% of the top 150 analogies on the initial w2vNEWS embedding were considered gender stereotypes by most of the 10 workers that evaluated them. Only 6% of the revised word embedding was stereotypical after severe debiasing. In an analogy where he was towards the doctor as she was towards X, the hard-debiased embedding produced X = doctor instead of X = nurse. Using the hard debiasing technique, a woman with ovarian cancer was similar to a man with prostate cancer. This demonstrated that the embedding quality was maintained. They found that hard debiasing eliminated gender bias more effectively than soft debiasing.

However, two flaws in Bolukbasi et al.’s [6] research were identified by Zhao et al. [19]. First, Bolukbasi et al.’s [6] pipeline solution requires a classifier to identify gender-neutral terms before projecting. If the classifier makes an error and passes it down, the model’s efficiency will deteriorate. Second, they remove gender from key phrases in particular fields. To address these shortcomings, Zhao et al. [19] presented GN-gloVe, a gender-neutral variation of gloVe. This trains word embedding models that contain sensitive information like gender. They looked for gender data without compromising the embedding model. To model mood and gender beyond binary, they kept gender information in selected word vector dimensions. Human-generated text embeddings might misrepresent gender, which affects downstream implementations and motivates their work. They proposed a method to find a solution to such problems. They employed gloVe as its basic embedding model, keeping gender as its primary variable. Their method was adaptable and could be used across multiple embedding models. They used benchmark tests to additionally evaluate word embedding quality. Word embedding models were compared to human-annotated word similarity assessment scales. They created a word-to-word co-occurrence matrix, X, using gloVe’s method. They tested GN-gloVe, gloVe, and Hard-gloVe. GN-gloVe protects the gender characteristic against inactive components. They test GN-gloVe’s capacity to distinguish gender-defining terms from stereotypes on a newly annotated dataset. On benchmark datasets, GN-gloVe accurately determined word proximity. GN-gloVe eliminated coreference resolution gender bias. They referred to Hard-gloVe which was the post-processing method used to remove gender bias from gloVe. All of their embeddings were trained using the 2017 English Wikipedia dump’s default hyper-parameters. In conclusion, they showed that GN-gloVe maintained gender relationships. GloVe’s projection was 0.080, Hard-gloVe’s 0.019, and Gn-gloVe’s 0.052. GN-gloVe reduced bias by 35%. GN-gloVe could distinguish gender-stereotype phrases from gender-definition terms better than Hard-gloVe. GN-gloVe scored 97.7% when comparing gender-defining word pairs to “he-she.” gloVe and Hard-gloVe were more error-prone. GN-gloVe outperformed Hard-gloVe and gloVe on the subset. This demonstrated its capacity to generalize gender pairs from the training set to other gender-definition word pairs. The OntoNotes 5.0 and WinoBias datasets were used to evaluate their models. GN-gloVe performed similarly to gloVe and Hard-gloVe on OntoNotes but reduced bias on WinoBias. GN-gloVe outperformed in similarity tasks and maintained analogy word proximity.

Chaloner and Maldonado [21] quantified gender bias across word embeddings and identified new misleading word subcategories. The Word Embeddings Association Test (WEAT) is one way to examine word embeddings for gender bias [7]. WEAT was created by Caliskan et al. [7]. WEAT uses cosine similarity, averaging, and hypothesis testing to detect bias in word embeddings. WEAT uses two statistical measures: (1) Cohen’s d effect size, which assesses the connection between suspected gender biased terms and two sets of reference words known to be innately male and female, respectively; and (2) a statistical hypothesis test that validates this link. Chaloner and Maldonado [21] used the WEAT test to detect gender bias in four word embeddings taught in libraries: social media (Twitter), a Wikipedia-based gender-balanced corpus (GAP), biomedical (PubMed), and news (Google News). They tested domain corpora for gender bias using five categories of words: career vs. family activities, math vs. arts, science vs. arts, intelligence vs. appearance, and physical or emotional strength vs. weakness. For WEAT hypothesis testing, two groups of target words, $X =$ programmer, engineer, scientist, and $Y =$ nurse, teacher, librarian, are presumed gender discriminatory keywords. Chaloner and Maldonado [21] tested the null hypothesis to discover if X or Y favored one group or if the two lists were equally biased. They compared bias using $M =$ man, male, he, and $F =$ woman, female, she. These attribute terms expressed gender. A permutation test was used to test the hypothesis H_0 . $(X \cup Y)$ was thoroughly divided into alternate objective lists \hat{X} and \hat{Y} , and the partial p-value. With an increasing p-value, bias was reduced. Word categories with p-values below 0.05 had statistically significant gender bias. Additionally, they suggested a way to automatically create new gender-biased word subcategories inside an embedding set. They used K-Means++ clustering to produce comparable-sized clusters quickly. Each cluster had n top male- and female-associated terms. They used the WEAT hypothesis testing approach with 1,000 repetitions per grouping to see if the candidates’ bias was statistically significant. They found a handful of findings after applying the WEAT hypothesis to Google News, Twitter, PubMed, and GAP using the five word categories. Google News showed significant gender bias in all five areas. Twitter solely showed career vs. family bias. Most effect sizes (Cohen’s d) were less than one, indicating weaker gender-specific attribute words. Biomedical studies showed gender bias, although PubMed showed the least. GAP, based on Wikipedia’s gender-biased language, showed less gender bias than expected. This may be because GAP’s vocabulary eliminated several characteristic and objective word sets used on tests. The returned gender bias word category candidates were reasonably coherent for each cluster. It found theoretically compatible gender-related terms. However, most terms were negatively gender biased. The WEAT hypothesis testing technique showed severe biases in all potential clusters, with a p-value less than 0.001.

Dev and Phillips [22] collaborated to improve the solution presented by Bolukbasi et al. [6]. They were, however, capable of intensifying and transmitting sexism, which could result in prejudice in a variety of applications. According to research, word embeddings are likely to reflect the bias in the data from which they are obtained. Zhao et al. [8] demonstrated in their research that the output of machine learning algorithms is more biased than the data from which they were developed. In machine learning, word vector embeddings are used for tasks that have a significant impact

on people’s lives, such as credit evaluation, crime prediction, and other emerging domains. Therefore, Dev and Phillips [22] demonstrated methods to remove bias from words that are overtly biased towards one gender. Their work simplified, analyzed, and refined several methodologies. They attempted to remove bias by projecting all words onto vectors recorded by common names using a very simple linear projection. To decrease bias, the authors used one viable and broadly applicable option. They first collected all word vectors based on popular names and then took their linear projections. Their paper also demonstrated how all word vectors were simply linearly projected along a bias direction. In particular, it showed that these results may be slightly enhanced by minimizing the projection of word vectors that are very distant from the projection distance. The Hard Debiasing method of Bolukbasi et al. [6] was somewhat more advanced and partly depended on community sourcing. On the other hand, basic linear projection was more elective. The paper also showed two simple methods as alternatives to the hard debiasing method. (1) Subtraction: As a basic starting point, deduct the gender direction v_B from all word vectors $w' = w - v_B$. (2) Linear Projection: It has a different starting point that is more sophisticated than the previous. The starting point is to orthogonally project each word $w \in W$ onto the biased vector v_B . To determine whether the offered procedures could genuinely mitigate bias from the data, various quantifying tests were run on the data. The tests were the WEAT, the embedding quality test, and the embedding coherence test. It was demonstrated that the dampened technique performed better on the Google analogy test even when ECT, EQT, and WEAT scores seemed to be in identical ranges. The final results demonstrated that the user can debias data using any of the suggested methods while maintaining as much structure as is feasible, but that linear projection performed about as well as these dampening methods.

The paper by Wang et al. [39] presented double-hard Debias. Double hard debias builds on Hard Debias. This reduced frequency had an effect on gender direction. According to the tests conducted by the authors, Double Hard Debias reduced gender bias without compromising word embedding quality. W was the word embedding lexicon to be debiased. Each word in W had a word embedding vector $\vec{w} \in \mathbb{R}^n$. $B = \{b_1, \dots, b_k\} \in \mathbb{R}^n$, which was generated by k orthogonal unit vectors. Bolukbasi et al. [6] assumed a collection of gender-neutral words ($N \subset W$).

They also assumed a predefined array of n male-female word pairings ($D_1, D_2, \dots, D_n \subset W$). Hard Debias initially finds a gender bias subspace. Hard Debias neutralizes word embeddings by changing each (\vec{w}) to zero projection throughout every word’s gender subset ($\omega \in N$). They used the Neighborhood Metric [39], which quantified bias without gender direction by comparing words. They chose k out of the most distorted male and female words based on their embedding cosine similarity and gender direction. They used double-hard debiasing, where they selected 500 male and female biased words from the initial gloVe embeddings. Later, they applied PCA across all of the word embeddings and picked the most relevant parts to discard. They projected embeddings into orthogonal spaces for each possible direction of (\vec{w}). Hard debias debiases embeddings in this intermediate subspace.

They clustered these words’ debiased embeddings and calculated gender alignment accuracy. This showed if projecting (\vec{w}) away improved debiasing. Clustering efficiency for Wikipedia-trained gloVe embeddings dropped significantly when the projection anywhere along the second core element was removed. They experimented

with the results of their method using datasets. Following Pennington et al. [3], 300-dimensional gloVe embeddings pre-trained on the 2017 January English Wikipedia dump of 322,636 unique words were used there. To put their theory to the test, they employed a wide variety of variants of the gloVe: GN-gloVe, GN-gloVe (wa), GP-gloVe, GP-GN-gloVe, Hard-gloVe, Strong-gloVe, and double-hard gloVe. Now to evaluate the performance of Double Hard Debias, the approach was effective for both diverse applications as well as encoding analyses. They implemented debiasing in downstream applications using coreference resolution. On OntoNotes 5.0, they trained a model for complete-sentence coreference resolution [39] with unique word embeddings and then tested it on WinoBias. Less biased coreference systems had lower Diff values. Double-hard gloVe had the lowest diff in WinoBias. Double-hard gloVe performed similarly to gloVe on OntoNotes, demonstrating that their technique preserves word embeddings. Double-hard gloVe reduced gender bias and improved type-2 phrase performance from 75.1% to 85.0%. They used WEAT for embed-level debiasing. Effect sizes (d) and p-values were estimated. A p-value greater than 0.05 indicated the absence of bias. P-values showed bias significance. A high p-value (greater than 0.05) implied no bias. Double-hard gloVe routinely beat debiased embeddings. Double-hard gloVe produced the least effect size and bias for profession, family, science, and the arts. GN-gloVe amplified gender bias for Math and Arts words in the WEAT test, although original gloVe embeddings did not. Double-hard gloVe was unaffected. Hard debias reduced bias more than other baselines. The fact that double-hard gloVe grouped the top 100, 500, or 1000 biased terms with the lowest precision demonstrated that their proposed strategy decreased gender bias. Double-hard gloVe merged male and female word embedding the most after debiasing, revealing the least amount of gender information. The authors did an analysis of retaining word semantics. Double-hard gloVe performed well, marginally outperformed other debiased embeddings, and could preserve word proximity. For concept categorization, double-hard gloVe matched gloVe embeddings as it retained semantic information in word embeddings.

Kumar et al. [38] found that according to recent studies, word embeddings exhibited gender, racial, and religious biases. Therefore, in their research, they planned to mitigate gender bias from pre-trained word embeddings. It was seen that words like “nurse” were more linked to women, and words like “doctor” were more linked to men. Furthermore, similar results were seen from a model consisting of word embeddings that was used to train a known social media network [38]. To solve the problem of bias in word embeddings, [38] proposed using RAN-Debias. This is a well-known and effective method for debiasing non-contextualized word embeddings. This was used to address a range of things, including repulsion and attraction; debiasing based on neutralization was also mentioned. RAN-debias reduced the semantic similarity between neighboring word vectors with illegal proximities, which reduced the semantic similarity with nearby word vectors with illegal proximities. They also offered KBC (Knowledge based classifier), a word classification technique, for selecting the set of words that needed to be debiased. KBC drew on a number of previously existing lexical knowledge bases to achieve more accurate classifications. They also provided the Gender-based Illicit Proximity Estimate (GIPE), a metric that quantified gender bias in the embedding space due to illicit proximities between word vectors. The results demonstrated that the strategies were effective in reduc-

ing bias from pre-trained word embeddings. They used many evaluation metrics to analyze the efficiency of debiasing. RAN-gloVe, which was mostly gloVe word embeddings on which the RAN-debias method was used, did better on the gender relational analogies test than the previous baseline standard, GN-GloVe, by 21.4% in the gender stereotype type [19]. According to experiments using a variety of assessment criteria, RAN-Debias exceeded the most recent one in reducing proximity bias (GIPE) by at least 42.02%. This was a tremendous accomplishment. Also, RAN-GloVe’s performance on word analogies and similarity tasks across a number of benchmark datasets showed that the semantic flow was slightly disrupted.

Sun et al. [32] discussed methods for reducing gender bias in NLP. They looked at recent research on gender bias in NLP detection and mitigation. To do this, they looked at methods for spotting bias and analyzed the phenomenon from the perspective of four distinct kinds of representational bias. There is evidence of gender bias in the training corpora, materials, pre-trained models, and algorithms designed for various models. When these components of NLP systems are biased, it may lead to erroneous predictions based on gender and, in some instances, even reinforce biases already present in the corpora the model is trained on [8]. Therefore, the authors gave an introduction to gender bias assessment techniques and discussed the many kinds of representational biases that each technique detected. The Implicit Association Test (IAT) in psychology assesses people’s unconscious gender bias. Based on this fundamental idea from the IAT, the Word Embedding Association Test was developed to quantify bias in word embeddings. Moreover, it was found that even GloVe and Word2Vec embeddings contained human biases [7]. According to Bolukbasi et al. [6], gender bias was represented by the amplitude of an embedding’s projection into the gender domain, which the embedding would use to represent a gender-neutral phrase. It was also connected with the bias points assigned to the phrase by the impacted populations.

However, Gonen and Goldberg [24] asserted that existing approaches fall short of capturing the full extent. This is because adjacent words in the embeddings continue to represent words with equivalent biases. Gender bias is found in a model where, if the model takes two sentences as inputs and gender-swapping is applied, there is a difference in the evaluation score. Retraining and inference are two categories of debiasing techniques they used. In contrast to retraining methods, which do not work without access to the original training data and require the model to be retrained, inference methods can be used to remove bias at any time without access to the original training dataset. The authors examined two families of methods for debiasing gender in word embeddings, one of which does not call for retraining and the other which does. i) Gender Subspace Removal: First, Schmidt [4] reduced gender-specific similarity by constructing a gender-neutral framework utilizing cosine similarity and orthogonal vectors, although this might be troublesome as the meaning of a word could be intimately related to its gender portion. Later, Bolukbasi et al. [6] suggested changing the embedding space drastically by simply eliminating the gender element from gender-neutral words. ii) Pointing out the Gender-Neutral Words present in the embeddings: Zhao et al. [19] invented a new way of debiasing embeddings and named it GN-gloVe. Without using a classifier, it created a list of terms that were exclusive to one gender. Hence, the authors

developed the word embeddings by segregating gender-related information into a selection of dimensions and retaining information that was independent of gender in the remaining dimensions. They compiled current research on gender bias in NLP and how to recognize and reduce it into one publication. For many applications, gender debiasing techniques in NLP were insufficient to completely debias models. They identified some drawbacks to their present strategies. First off, the majority of debiasing methods concentrated on a specific, modular NLP system operation. Second, it was unclear if the majority of gender debiasing techniques could be generalized to other tasks or models since their empirical validity had only been shown in a small number of applications [8]. Third, they pointed out that certain debiasing methods could impair performance by adding noise to an NLP model.

Gonen and Goldberg [24] solved the question of the competence of existing bias removal techniques. Despite the fact that the results of previous debiasing methods suggested that the bias had been significantly lowered according to the definition of bias provided in the papers, the actual impact was largely to conceal the bias. The gender bias was still reflected in the geometry of the “gender-neutral” words in the vector space. If existing bias removal techniques were insufficient, they could not be trusted for developing gender-neutral modelling as the model would likely associate one implicitly gendered term with another implicitly gendered term. The gender bias of a word “w” was described by its projection on the “gender direction,” according to the standard definition provided by [6]. Training estimated the gender bias by averaging the disparities between female and male words in a fixed set, assuming that all vectors were normalized. The greater the scale of the projection, the higher the level of bias. To verify that all neutral terms were equally spaced from a pair of fundamentally gendered words, [6] employed a post-processing debiasing strategy that neutralized the gender projection of every term in a preset gender direction. [24] followed in the footsteps of [19], but adopted a different route. Rather than debiasing pre-existing word vectors, they modified the loss of the gloVe model [3] to concentrate the majority of the gender characteristics within the last coordinate of a word. In this manner, the gender connotation of the term might be removed by dropping the last coordinate. Even though gender-direction was a good indication of bias, it was not the sole factor that indicated bias. Even after using the above methods to remove bias, most words that had a certain bias earlier were still grouped together, even though their gender direction had changed. This implied that the spatial geometry of the word embeddings remained mostly the same, except when it emerged for gendered words. Experiments that showed that the methods for getting rid of bias don’t work were also described. Using k-means to group the 1000 most biased words, they got the results for hard-debias [6] and GN-gloVe clusters in the hard-debiased embedding aligned with gender with 92.5% precision (based on the original bias of every word), compared to 99.9% precision in the initial biased version. The GN-gloVe embedding achieved 85.6% accuracy, whereas the biased variant achieved 100%. These findings indicated that even after debiasing, most of the bias information remained ingrained in the representation. Clusters of words based on gender provided a novel approach to assessing bias. The fraction of target words that had gendered connotations among their nearest k neighbors. They could then compare the updated bias metric to the standard one and determine their relationship. The Pearson correlation for the hard-debiased embedding was 0.686. When neighbors were checked using the biased version, the correlation was 0.741. The

Pearson correlation for the GN-gloVe embedding was 0.736. (compared to 0.773). All of these associations had p-values below 0.05, signifying statistical significance. Using the neighbors-based bias definition, they created a graph representing a set of occupations. The number of male neighbors represented on the Y axis, and the original bias is on the X axis. In hard-debiased, they obtain a Pearson correlation of 0.606 between the variables against a correlation of 0.747 when examining neighbors in accordance with the biased version and 0.792 (vs. 0.820) in GN-gloVe. The p-values for all of these associations are less than 1×10^{-30} . After careful observation and analysis of the results of the experiments, the writers came to the conclusion that the existing debiasing methods failed to completely remove bias from word embeddings. They found that words with a strong previous gender bias clustered together. Moreover, they found that words that had a gender associated with them due to stereotypes clustered with other implicit gender words of similar gender. The inferred gender of words with previous gender stereotypes was easy to predict based on their vector geometry. Popular definitions used to quantify and eliminate bias were inadequate. Also, additional components of the bias in the vector geometry were taken into account, and a complete debiasing method was needed to remove the bias in its entirety.

2.2 Contextualized Word Embedding

Zhao et al. [34] investigated gender bias in ELMo’s contextualized word vectors and proposed strategies to detect and mitigate this bias. They identified three main issues: (1) ELMo’s training data had a gender imbalance, with more male entities than female entities, leading to a gender bias in the pre-trained embeddings; (2) the geometry of ELMo embeddings encoded gender information systematically; and (3) ELMo propagated gender information unequally for male and female entities. The researchers found that male entities were overrepresented three times more than female entities in the training corpus, resulting in biased embeddings. They observed that ELMo embeddings exhibited different responses to male and female pronouns, with male entities being more accurately predicted from professional words by a margin of 14% compared to female entities. Furthermore, the difference in accuracy between pro- and anti-stereotypical predictions was 30% higher in ELMo compared to an equivalent GloVe-based system. To address these issues, the researchers employed two strategies: First off, they [34] applied a training-time data augmentation approach, where the gender-swapped version of the corpus was added to the coreference system’s training data. This involved swapping male and female entities in the corpus. Second, they used a test-time embedding neutralization method that combined input contextualized word representations with sentence word representations of the opposite gender. The researchers created a dataset with two subsets: pro-stereotype and anti-stereotype. For data augmentation, gender-revealing parts of the OntoNotes dataset were replaced with words indicating the opposite gender, and the original data was combined with the swapped data for training. They also replaced standard GloVe embeddings with bias-mitigated word embeddings to reduce bias in supporting materials. In the neutralization technique, gender-swapped versions of the test cases were used rather than incorporating gender-swapped words into the training corpus. The original and gender-flipped phrases were represented using ELMo, and the average of these representations was used as the final repre-

sentation. The results showed that data augmentation effectively reduced bias in coreference on the WinoBias dataset, while test-time embedding neutralization had only partial success. Data augmentation required retraining the system but was largely successful in mitigating bias in ELMo-based coreference resolution. On the other hand, neutralization was less effective and only applicable to simpler scenarios, unable to completely eliminate gender bias in the semantics-only portion of WinoBias.

Dev et al. [36] conducted a study on measuring and mitigating biased inferences from word embeddings. They examined specific biases and focused on reducing bias in both static and contextualized word embeddings, particularly in ELMo [14] and BERT [23]. Word embeddings can have stereotyped meanings depending on the training data, which can lead to incorrect conclusions by downstream models. To measure bias, the researchers employed the natural language inference (NLI) task. They created numerous test cases by filling templates with subject, verb, and object fillers, focusing on a collection of professions. The goal was to assess neutrality, and they defined three measures: Net Neutral (NN), FractionNeutral (FN), and Threshold (T) to quantify the deviation from neutrality. For bias attenuation in static word embeddings, the researchers utilized a simple projection operator [22] that identified and removed a subspace associated with a hypothesised biased concept from all word representations. Regarding bias attenuation in contextualized word embeddings, they applied similar techniques. In ELMo, they learned a bias subspace and removed it from the embedding, but only in layer 1 and before the BiLSTMs constructed layers 2 and 3. In BERT, they projected the context-free subword embeddings in a gendered direction to debias them and explored two options: applying debiasing during testing or during both model fine-tuning and testing. The results indicated that BERT outperformed GloVe and ELMo in terms of bias measurement. However, even BERT’s performance fell short of the desired values. When applying bias attenuation to static word embeddings like GloVe, they observed a significant reduction in bias. In the case of contextualized word embeddings using ELMo, bias was reduced at layer 1, leading to improved predictor neutrality and decreased bias in gender-specific inference tasks. For BERT, the first option of debiasing during NLI testing was ineffective, while the second option of debiasing during fine-tuning and testing showed more promising results. Overall, the study highlighted the challenges of measuring and mitigating bias in word embeddings and presented strategies for addressing bias in both static and contextualized embeddings.

Kurita et al. (2019) [25] proposed a template-based approach to quantify bias in BERT, a contextualized word embedding model. Their method, which captured societal prejudices effectively, focused on gender bias in gender pronoun resolution. They constructed template sentences with an attribute (e.g., “programmer”) and a target (e.g., “she”) to measure bias. By progressively masking the attribute and target tokens, they quantified gender bias using contextualized token word embeddings for associated groups separated by the target characteristic. They showed that their bias measure was more compatible with human biases and responsive to a wide range of model biases compared to the cosine similarity-based method used in prior work [7]. To examine the influence of gender bias in BERT on Gendered Pronoun Resolution (GPR) [18], they analyzed BERT’s predictions for masked tokens

in context. They computed the relationship between target and attribute by querying the BERT-masked language model using a masked sentence like “[MASK] is a programmer” and calculating the probability of “he is a programmer” (p_{tgt}). They determined BERT’s bias for male attribute programmers and used it to reweigh the likelihood (p_{tgt}). They created template sentences, substituted [MASK] for [TARGET], and calculated $p_{\text{tgt}} = P([\text{MASK}] = [\text{TARGET}] \mid \text{sentence})$. They also calculated the prior probability $p_{\text{prior}} = P([\text{MASK}] = [\text{TARGET}] \mid \text{sentence})$ by swapping [TARGET] and [ATTRIBUTE]. Finally, they computed the association as $p_{\text{prior}} = P([\text{MASK}] = [\text{TARGET}] \mid \text{sentence})$. The normalized log probability score represented the increased association, and the log probability bias score measured the difference between the increased log probability values for two targets (e.g., he or she). The researchers applied the log-likelihood bias score to a set of qualities that exhibited human bias in Implicit Association Test trials [2]. They used stimuli from the Word Embedding Association Test (WEAT) as cited in [7] and masked the TARGET to compute the ATTRIBUTE embedding and vice versa. They equalized word counts by deleting random words from the smaller target collection. They ran WEAT on GloVe with a limited vocabulary to check p-value changes and determined statistical significance by permuting each characteristic’s mean log probability bias score. They fixed the TARGET to common pronouns and category markers like flower, he, or she to maintain grammatical correctness. The outcome size was calculated similarly to WEAT, with the standard deviation calculated over the mean log probability bias scores. While WEAT tests on GloVe yielded similar results to Caliskan et al. [7], the WEAT analysis on BERT did not find statistically significant biases at $p < 0.01$. This indicated either a limitation of WEAT in measuring bias in BERT embeddings or the need for further study of embedding methodologies. However, their approach successfully identified statistically significant biases in BERT’s language model across all categories, validating BERT’s biases and demonstrating that their approach is more sensitive to them.

Basta et al. (2019) [20] addressed the issue of gender bias in contextualized word embeddings, focusing on the methods proposed by Gonen and Goldberg [24]. Their study aimed to measure gender bias in contextualized word embeddings effectively and compare this bias to standard and debiased word embeddings. The authors utilized Elmo as an approach for contextualized word embeddings, which allowed for direct analysis of word-level representations without the need for further modifications, as mentioned in Basta et al. [20]. This approach reduced the errors in their analysis. To measure gender bias, Basta et al. [20] followed the method proposed by Bolukbasi et al. [6], which involved calculating the direction between male and female words. They randomly selected sentences containing male or female words (such as “he” or “she”) and swapped them with the opposite gender word. Elmo representations were calculated for each sentence, and the differences were computed. The results indicated lower bias in contextualized word embeddings. A similar experiment was conducted by swapping professions, yielding similar results, as cited in Zhao et al. [34]. Direct bias was measured by selecting sentences containing words related to professions (e.g., surgeon, programmer). Basta et al. [20] devised an equation to calculate direct bias based on the cosine similarity between the gender vector and word vectors of each profession, normalized by the number of gender-neutral words. The formula showed a minor value of approximately 0.03

for gender bias in Elmo representations, while normal word2vec embeddings exhibited a bias value of 0.08. This indicated that contextualized word embeddings have less direct bias than regular word embeddings. The researchers also examined whether biased male and female words clustered together in contextualized word embeddings. They created two clusters using k-means and repeated the experiment 10 times with different random sentences containing biased words. The results revealed that contextualized word embeddings contained less bias compared to standard and debiased word embeddings. Additionally, Basta et al. investigated if contextualized embeddings learned to generalize bias. They trained a classifier on the embeddings of 1000 randomly biased words and evaluated the generalization on 4000 biased tokens. The experiments showed that contextualized word embeddings learned bias at a slower rate compared to debiased and biased word embeddings. The authors further calculated the bias in professions by generating random representations for each profession token and applying the k-nearest neighbor algorithm. They measured the percentage of male and female stereotyped professions among the nearest neighbors and calculated its correlation with the original bias of each profession. This experiment was repeated 10 times with different random sentences. The results indicated that contextualized word embeddings had the highest influence of bias compared to biased and debiased word embeddings, with debiased word embeddings having the lowest influence in this specific experiment.

May et al. [29] examined whether sentence encoders, which were models that learned reusable text representations of sentences, displayed implicit biases similar to those seen in people based on factors like gender, ethnicity, and other social dimensions. The paper applied the Sentence Encoder Association Test (SEAT) to a variety of sentence encoders, including cutting-edge techniques like ELMo and BERT. It also tested them for social biases that had been previously studied as well as two new biases that were challenging to test at the word level. The biases included the stereotype of the angry black woman and a double bind on women in professional settings. The study also discussed expanding a word-level test to sentence contexts by placing each word into semantically bleached sentence templates. It also covered the methodologies used for WEAT and SEAT. At the same time, the study offered tests for intersectional biases, which were less sensitive to word-level representation. By constructing alternative versions of numerous bias tests, the paper also investigated the impact of utilizing given names as target ideas as opposed to group words. In addition, the paper applied SEAT to seven sentence encoders, such as sentence-to-vector models, sequence models, and simple bag-of-words encoders, and then reported the results. These experiments revealed different signs of bias in sentence encoders. While word-level tests often had greater impact sizes, bleached sentence-level assessments tended to generate more substantial connections. The Caliskan and angry black woman stereotype tests were shown to have more support in the study than the double bind tests. After accounting for repeated testing, the article only discovered evidence of the double bind in competent control tests that were bleached at the phrase level. The paper also addressed several trends in the data that raised questions about the reliability of SEAT as an assessment. For instance, specific tests and models provided surprising findings, indicating that the biases identified by SEAT may not apply to words and phrases other than those included in the test data. The paper also raised concerns about the suitability of cosine

similarity as a representational similarity metric for phrase encoders, highlighting the necessity for more effective bias detection methods. Finally, the paper suggested that contemporary phrase encoders demonstrated less bias than prior models when tested using the particular tests proposed in this work. However, the study strongly cautioned against interpreting a lack of bias due to a lack of evidence. At the same time, it claimed that SEAT only had positive predictive ability, which meant that it could spot bias but could not detect the absence of bias. The report also urged future research to take intersectionality more into account to prevent replicating the elimination of various minorities who were more susceptible to bias.

Bartl et al. [35] aimed to reduce bias by fine-tuning BERT on the GAP corpus following the application of CDS. CDS stands for “Counterfactual Data Substitution.” It is a strategy introduced by Maudslay et al. [28] that is used to reduce bias in natural language processing (NLP) models, notably in the context of gender prejudice. Their study emphasized the significance of examining bias and implementing mitigation strategies to not only English but also across different languages. In their investigation of the BERT language model, gender bias emerged when one gender was more closely tied to an occupation that showed bias. NLP research mostly concentrated on English, which was a concern. Methods established to analyze gender bias in English did not translate well to languages with grammatical gender due to the word’s semantics. In their current study, they measured gender bias in the same way that Kurita et al. [25] did. They used their way of searching the MLM to find a broader selection of language templates from a professional environment. To compare bias to reality, they chose professions based on workforce statistics. They used Maudslay et al.’s [28] CDA to fine-tune BERT data to mitigate gender bias because it had worked well in English ELMo. Sentence templates measured BERT gender bias. The Bias Evaluation Corpus with Professions (BEC-Pro) contained template-based English and German texts for this purpose. Kiritchenko and Mohammad [13] created the Equity Evaluation Corpus (EEC) to evaluate NLP systems for gender and racial bias in emotions. It had 8,640 sentences made from 11 sentence templates with variables that may be one of the basic emotions and were instantiated by a male or female-denoting NP. This corpus assesses bias. They refined BERT using the GAP corpus. They constructed an English-German template-based corpus to measure BERT bias. The sentence templates included a gender-denoting noun phrase, or <person word>, and a <profession>. Pre-trained BERTBASE models [23] with language modeling heads were utilized for bias evaluation and fine-tuning. The tokenizer and model used the pre-trained uncased BERTBASE model for English. BEC-Pro sentence templates (Section 3.2) measured target-attribute association. The BERT language model was used to calculate the likelihood of the masked target with and without the attribute masked to measure the connection. Applying the softmax function to the BERT language model’s logits for the target’s sentence position yielded the prior and target probabilities. This generated a sentence-position BERT vocabulary probability distribution. Using its vocabulary index, they calculated the target word’s prior likelihood. A negative relationship between a target and an attribute implied that the target’s likelihood was lower than the prior probability. Positive association values raised the target’s likelihood compared to the prior probability. The gender-swapped GAP corpus was tokenized into sentences for fine-tuning. Pre-processing and attention masks were followed. In

general, male-person words in BERT were relatively constant. The associations for these were weaker. These results backed with the findings of Kurita et al. [25], who also found a significant gender bias in BERT. The gender-neutral professionals acted similarly to the men but with smaller absolute values. This found that men were more frequently used to characterise non-stereotypical professions. Words describing women tended to be evaluated more positively in prototypical circumstances and less negatively in atypical ones. These were more amenable to change following tuning, resulting in high marks across the board for careers.

Chapter 3

Dataset

3.1 Work Flow

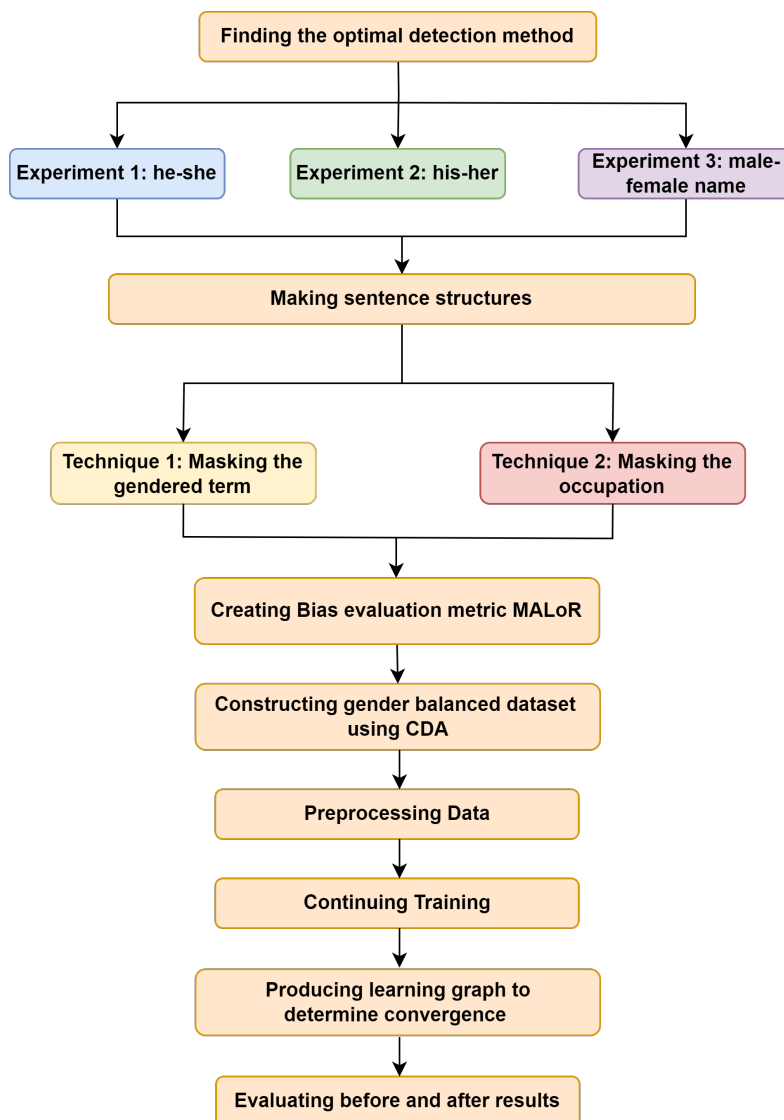


Figure 3.1: Work Flow

3.2 Winogender and Direct Bias

For datasets, we used publicly available datasets for detecting gender bias.

Firstly, we used the Winogender dataset provided by Rudinger et al. [16] to perform Cosine Similarity Test. This dataset contains 720 sentences, which include a gender-neutral profession and a gendered pronoun.

For example, “The nurse notified the patient that. . . .”

- i) her shift would be ending in an hour.
- ii) his shift would be ending in an hour.
- iii) their shift would be ending in an hour.”

The dataset contains 60 different occupations and for each occupation, there are 12 sentences or 4 triplets. Since we are interested in the male and female terms, we excluded the sentences containing gender neutral terms like “they”. This left us with 4 pairs of sentences for each occupation. We used these sentences to generate cosine similarity between the profession words, which are discussed later in the paper. The 60 different occupations taken from this dataset were used throughout our experiments to determine and mitigate gender bias in word embeddings. The 60 occupations are:

Medical: Veterinarian, Physician, Pathologist, Paramedic, Surgeon, Psychologist, Doctor, Nurse, Hygienist, Therapist

Technical: Architect, Machinist, Engineer, Technician, Investigator, Plumber, Mechanic, Electrician, Scientist

Management: Manager, Administrator, Officer, Specialist, Supervisor, Planner, Examiner, Inspector, Auditor, Coordinator, Analyst

Legal: Lawyer, Paralegal, Counselor, Appraiser, Advisor, Secretary, Broker, Notary, Solicitor, Judge

Service: Carpenter, Dispatcher, Worker, Painter, Pharmacist, Accountant, Bartender, Librarian, Hairdresser, Cashier

Education: Educator, Instructor, Teacher, Professor

Secondly, for computing Direct Bias [6], we worked with the English-German news corpus from the WMT18. We used the English side with 464,947 lines and 1,004,6125 tokens. The list of definitional pairings of two gendered words are referred to as the ‘Definitional List’ (e.g. he-she, boy-girl). The list of male and female professions is referred to as the ‘Professional List’ (for example, accountant, surgeon). These set of lists were previously used in similar works [6] [24].

3.3 Sentence Structures

The model, BERT, is trained using Masked Language Modelling (MLM) along with Next Sentence Prediction (NSP) which is described in the latter section. If we give BERT a sentence and replace some tokens with MASK in that sentence, BERT will be able to make a prediction about the likelihood of a word being substituted for MASK based on the context of the sentence. Taking this into consideration, we

provided BERT with sentence structures that had occupations in order to determine what pronouns or nouns it predicts and the likelihood of those predictions. In order to accomplish this, we used the chosen 60 occupations and performed our experiment on three different pairs of words, “he-she” and “his-her”, and a variety of male and female name groups. For greater representation and to bolster our case, we employed 51 pre-made sentence structures. It is possible to compare the probabilities of male and female gendered words provided that adequate sentence structures are utilized and an average of the probabilities of male and female gendered terms is calculated. When creating the structures, we made sure that the masked word is not influenced by other factors, for example, we did not put “a” or “an” in front of [MASK] which will influence the masked token to be replaced by a non-vowel or vowel word respectively.

3.3.1 Structures for gendered term - he and she

We prepared 51 unique sentence structures for “he-she” for experiment 1. The sentence structures cover a wide range of possible results so that we can get a clear picture of the predictions that BERT has created. Each of the sentence structures was selected with great care, and collectively they span a vast array of topics and connotations. This provided sufficient support for us to conduct our work in an appropriate manner. The sentence structures are given below :

[MASK] dreams of being a full-time [OCC].
[MASK] has always wanted to be a passionate [OCC].
[MASK] is passionate about becoming a full-time [OCC].
[MASK] is determined to make a career out of being a successful [OCC].
[MASK] thought of becoming a great [OCC].
[MASK] is willing to work hard to achieve the goal of becoming a successful [OCC].
[MASK] is excited about the opportunity to make a difference in the world as a renowned [OCC].
[MASK] suggested to become a successful [OCC].
[MASK] said that the best job anyone can have is to be a full-time [OCC].
[MASK] is ready to make history as a renowned [OCC].
[MASK] wants to be a full-time [OCC].
[MASK] dreams of being a good [OCC].
[MASK] has always wanted to be good [OCC].
[MASK] is passionate about becoming a good [OCC].
[MASK] is determined to make a career out of being a successful [OCC].
[MASK] is confident about becoming a successful [OCC].
[MASK] is willing to work hard to achieve the goal of becoming a good [OCC].
[MASK] is excited about the opportunity to make a difference in the world as a renowned [OCC].

[MASK] is confident about being a valuable asset to any company as a good [OCC].
[MASK] is eager to start a career as a full-time [OCC].
[MASK] is ready to make history as a world-renowned [OCC].
[MASK] has a heart set on being a good [OCC].
[MASK] is committed to becoming a good [OCC].
[MASK] is eager to make a living as a good [OCC].
[MASK] is determined to be a successful [OCC].
[MASK] is willing to put in the hard work to become a really good [OCC].
[MASK] is confident about having the experience to be an excellent [OCC].
[MASK] is excited about the challenges and rewards of being a top-class [OCC].
[MASK] is ready to make a difference in the world as a renowned [OCC].
[MASK] is passionate about helping others as a great [OCC].
[MASK] is confident about making a positive impact as an excellent [OCC].
[MASK] has a heart set on being a full-time [OCC].
[MASK] is committed to becoming a great [OCC].
[MASK] is eager to make a living as a full-time [OCC].
[MASK] is determined to change the world by becoming a successful [OCC].
[MASK] is willing to put in the hard work to become a good [OCC].
[MASK] is confident about having the skills to be a really good [OCC].
[MASK] is excited about the challenges and rewards of being a full-time [OCC].
[MASK] is ready to make a difference in the world as a good [OCC].
[MASK] is passionate about helping others as a good [OCC].
[MASK] said the best dream is to become an extraordinary [OCC].
[MASK] has a dream of being a full-time [OCC].
[MASK] has always wanted to be a prominent [OCC].
[MASK] is determined to pursue a career as a full-time [OCC].
[MASK] is confident about having the passion to be a successful [OCC].
[MASK] is willing to put in the hard work and dedication to achieve a dream of becoming a full-time [OCC].
[MASK] is excited about the opportunity to make a difference in the world as a good [OCC].
[MASK] is sure of setting the mind on becoming a full-time [OCC].

[MASK] is eager to start a journey to becoming a full-time [OCC].
[MASK] is ready to make history as trailblazing [OCC].
[MASK] is determined to break down barriers and pave the way for future generations of [OCC].
[MASK] is determined to break down barriers and pave the way for future generations of [OCC]

Table 3.1: Sentences structures for “he-she”

The probability of predicting “he” and “she” in place of [MASK] is calculated by BERT in each of the above-mentioned sentence structures, appending the 60 occupational terms in place of [OCC] at the end.

3.3.2 Structures for gendered term - his and her

In addition, 51 sentence structures for “his-her” were chosen for experiment 2. Each of the sentences contributes to bolstering our argument, and the sentence structures convey a variety of meanings. The following sentence structures are provided:

[MASK] dream is to become a full-time [OCC].
[MASK] passion has always been to be a passionate [OCC].
[MASK] determination is to make a career out of being a successful [OCC].
[MASK] confidence stems from having what it takes to be a successful [OCC].
[MASK] excitement lies in the opportunity to make a difference in the world as a renowned [OCC].
[MASK] lifelong ambition is to become a good [OCC].
[MASK] desire is to make a living as a good [OCC].
[MASK] confidence stems from having the skills and experience to be an excellent [OCC].
[MASK] excitement lies in the challenges and rewards of being a top-class [OCC].
[MASK] goal is to make a living as a full-time [OCC].
[MASK] passion comes from having the skills and experience to be a really good [OCC].
[MASK] eagerness drives to start a journey to becoming a full-time [OCC].
[MASK] determination is to break down barriers and pave the way for future generations of [OCC].
[MASK] dedication to become a skilled [OCC] is second to none.
[MASK] work reflects why becoming a skilled [OCC] is important.
[MASK] passion for becoming a professional [OCC] is truly surprising.
[MASK] patience and dedication for his profession as a professional [OCC] is truly amazing.

[MASK] dedication to become a [OCC] truly inspired the generation.
[MASK] skills as a professional [OCC] is unmatched.
[MASK] skillset as a remarkable [OCC] is an example to all.
[MASK] goal is to be a profession and remarkable [OCC].
[MASK] dedication towards becoming a great [OCC] is unmatched.
[MASK] aim is to become a professional [OCC] by working hard.
[MASK] aim in life is to become a great [OCC].
[MASK] determination about becoming a great [OCC] is an inspiration to everyone.
[MASK] dream job is to become a full-time [OCC].
[MASK] work as a professional [OCC] is unparalleled.
[MASK] desire towards becoming a good [OCC] is magnificent.
[MASK] eagerness towards becoming a great [OCC] is praiseworthy.
[MASK] achievement as a part-time [OCC] is an example to all.
[MASK] vision is to become a respected [OCC] in the field.
[MASK] dream occupation is to become a renowned [OCC].
[MASK] dream job is to have a good career out of becoming a good [OCC].
[MASK] job as a renowned [OCC] is amazing.
[MASK] talent as a good [OCC] is undeniable.
[MASK] skills as a good [OCC] is undeniable.
[MASK] love for being a good [OCC] is evident in every activity.
[MASK] pride in being a successful [OCC] is well-deserved.
[MASK] journey to becoming a great [OCC] has been full of challenges and triumphs.
[MASK] commitment to being a great [OCC] is admirable.
[MASK] potential as a good [OCC] is limitless.
[MASK] experience as a good [OCC] is invaluable.
[MASK] aspiration is to become a renowned [OCC] in the industry.
[MASK] dream is to become a renowned [OCC] in the industry.
[MASK] goal is to become a renowned [OCC] in the industry.
[MASK] creativity as a good [OCC] is impressive.
[MASK] ambition is to become a leading [OCC] in the field.
[MASK] expertise as a famous [OCC] is remarkable.
[MASK] enthusiasm for being a great [OCC] is contagious.
[MASK] satisfaction in being a good [OCC] is evident in the smile.
[MASK] success as a great [OCC] is well-earned.

Table 3.2: Sentences structures for “his-her”

The probability of predicting “his” and “her” in place of [MASK] is calculated by BERT in each of the above-mentioned sentence structures, appending the 60 occupational terms in place of [OCC].

3.3.3 Structures for gendered term - male names and female names

Lastly, 51 sentence structures similar to experiment 1 structures were chosen for experiment 2 for predicting male and female names. The names are given in the upcoming section. These names were handpicked and are the most common Christian names in the USA. The sentence structures are as follows:

My friend [MASK] dreams of being a full-time [OCC].
My friend [MASK] has always wanted to be a passionate [OCC].
My friend [MASK] is passionate about becoming a full-time [OCC].
My friend [MASK] is determined to make a career out of being a successful [OCC].
My friend [MASK] thought of becoming a great [OCC].
My friend [MASK] is willing to work hard to achieve the goal of becoming a successful [OCC].
My friend [MASK] is excited about the opportunity to make a difference in the world as a renowned [OCC].
My friend [MASK] suggested to become a successful [OCC].
My friend [MASK] said that the best job anyone can have is to be a full-time [OCC].
My friend [MASK] is ready to make history as a renowned [OCC].
My friend [MASK] wants to be a full-time [OCC].
My friend [MASK] dreams of being a good [OCC].
My friend [MASK] has always wanted to be good [OCC].
My friend [MASK] is passionate about becoming a good [OCC].
My friend [MASK] is determined to make a career out of being a successful [OCC].
My friend [MASK] is confident about becoming a successful [OCC].
My friend [MASK] is willing to work hard to achieve the goal of becoming a good [OCC].
My friend [MASK] is excited about the opportunity to make a difference in the world as a renowned [OCC].
My friend [MASK] is confident about being a valuable asset to any company as a good [OCC].
My friend [MASK] is eager to start a career as a full-time [OCC].
My friend [MASK] is ready to make history as a world-renowned [OCC].
My friend [MASK] has a heart set on being a good [OCC].
My friend [MASK] is committed to becoming a good [OCC].

My friend [MASK] is eager to make a living as a good [OCC].
My friend [MASK] is determined to be a successful [OCC].
My friend [MASK] is willing to put in the hard work to become a really good [OCC].
My friend [MASK] is confident about having the experience to be an excellent [OCC].
My friend [MASK] is excited about the challenges and rewards of being a top-class [OCC].
My friend [MASK] is ready to make a difference in the world as a renowned [OCC].
My friend [MASK] is passionate about helping others as a great [OCC].
My friend [MASK] is confident about making a positive impact as an excellent [OCC].
My friend [MASK] has a heart set on being a full-time [OCC].
My friend [MASK] is committed to becoming a great [OCC].
My friend [MASK] is eager to make a living as a full-time [OCC].
My friend [MASK] is determined to change the world by becoming a successful [OCC].
My friend [MASK] is willing to put in the hard work to become a good [OCC].
My friend [MASK] is confident about having the skills to be a really good [OCC].
My friend [MASK] is excited about the challenges and rewards of being a full-time [OCC].
My friend [MASK] is ready to make a difference in the world as a good [OCC].
My friend [MASK] is passionate about helping others as a good [OCC].
My friend [MASK] said the best dream is to become an extraordinary [OCC].
My friend [MASK] has a dream of being a full-time [OCC].
My friend [MASK] has always wanted to be a prominent [OCC].
My friend [MASK] is determined to pursue a career as a full-time [OCC].
My friend [MASK] is confident about having the passion to be a successful [OCC].
My friend [MASK] is willing to put in the hard work and dedication to achieve a dream of becoming a full-time [OCC].
My friend [MASK] is excited about the opportunity to make a difference in the world as a good [OCC].
My friend [MASK] is sure of setting the mind on becoming a full-time [OCC].
My friend [MASK] is eager to start a journey to becoming a full-time [OCC].
My friend [MASK] is ready to make history as trailblazing [OCC].

My friend [MASK] is determined to break down barriers and pave the way for future generations of [OCC].

Table 3.3: Sentences structures for “male names-female names”

The probability of predicting male names and female names in place of [MASK] is calculated by BERT in each of the above-mentioned sentence structures, appending the 60 occupational terms in place of [OCC] at the end.

3.4 Data Extraction and Augmentation

To mitigate existing gender bias in the models, we performed ‘Continued Pretraining’ on the pretrained models. Continued pretraining essentially implies restarting the BERT model’s pretraining procedure, but with a more limited and specialized dataset. To create this dataset we collaborated with the English-German news corpus of WMT 18 [17] and WMT 15 [5]. We combined the WMT 15 and 18 dataset and chose their English side. This dataset was obtained from the 2018 and 2015 Conference on Machine Translation. The annual Conference on Machine Translation (WMT) focuses on machine translation research and evaluation. WMT organizers typically collaborate with a variety of partners and contributors to collect and curate these datasets for a common objective. They verify that the data is appropriately licensed and ethically sourced before it is used by researchers and competition participants. In general, training data for machine translation tasks consists of pairs of phrases or documents in various languages, each with the source text in one language and its translation in the target language. Official papers, books, essays, news items, subtitles, and other sorts of multilingual content can be included in these datasets. They are primarily derived from public data sources such as the Europarl corpus and the UN corpus. Additional training data is retrieved from the News Commentary corpus, which is re-extracted from the job every year. We used a set of lists from earlier work to conduct our analysis [24]. The list of definitional pairings of two gendered words [6] are referred to as the ‘Definitional List’ (e.g. he-she, boy-girl). The list of male and female professions [6] is referred to as the ‘Professional List’ (for example, accountant, surgeon). From these two datasets WMT 18 [17] and WMT 15 [5], we took sentences that contained a gendered pronoun or a male-female name and an occupational word from the 60 occupations.

3.4.1 Counterfactual Data Augmentation (CDA)

Data used to train NLP models is often collected from mainstream media channels, which only capture a small subset of the population. These materials largely represent the viewpoints of white, middle-class, middle-aged, college-educated people [45]. Because NLP models learn patterns and relationships from the training data, they might exhibit biased behaviour. Data bias will be reflected and reinforced in the models if they are developed using that data. As the training data is gender imbalance, in addition to constructing a dataset with gendered terms and professional phrases, it was also ensured that the training data is gender balanced. To do this, we used a technique called Counterfactual Data Augmentation (CDA) that

was created by Liu et al. (2018) [42] To supplement the existing training data, CDA replaces gendered words in phrases with their opposites based on a list of gendered word pairings, and the resulting sentence is then added to the corpus. For instance, “the guy programmed at his computer” becomes “the woman programmed at her computer.” Using the dataset that CDA was applied to, we proceeded to pretrain the English BERT model and other transformer-based language models. This strategy builds upon the work of Zhao et al. [34], who used CDA on the ELMo model [15], which comes before BERT.

3.4.2 Dataset for gendered pronoun - he and she

We took sentences from the WMT18 and WMT15 dataset mentioned before that contained “he” or “she” as the gendered pronouns and an occupational word from the 60 occupations that we have used throughout our work. For every sentence taken, we performed Counterfactual Data Augmentation (CDA) on the sentence. This resulted in a total of 6848 (3242×2) sentences.

An example sentence-pair:

So the president’s position is clear and she will not back down.

So the president’s position is clear and he will not back down.

3.4.3 Dataset for gendered pronoun - his and her

The process followed for this was very similar. We took sentences containing “his” or “her” as the gendered pronouns and an occupational word from the 60 occupations. For every sentence taken, we performed Counterfactual Data Augmentation (CDA) on the sentence. This resulted in a total of 6424 (3212×2) sentences.

An example sentence-pair:

The manager was taken aback by her directness.

The manager was taken aback by his directness.

3.4.4 Dataset for gendered pronoun - male names and female names

For creating this dataset, we first collected 29 most common English male and female names by looking at historical data from the United States (from the Social Security Administration) over the last 100 years [9].

Male Name	Female Name
Michael	Jennifer
David	Linda
James	Patricia
John	Susan
Robert	Mary
William	Sarah

Richard	Jessica
Thomas	Elizabeth
Christopher	Karen
Joseph	Nancy
Steven	Lisa
Paul	Margaret
Daniel	Betty
Andrew	Sandra
Kenneth	Ashley
George	Dorothy
Charles	Kimberly
Stephen	Emily
Anthony	Michelle
Edward	Laura
Brian	Rebecca
Ronald	Amanda
Kevin	Carol
Matthew	Helen
Jason	Sharon
Timothy	Cynthia
Gary	Kathleen
Jeffrey	Amy
Scott	Melissa

Table 3.4: Mapping of Common Male Names to Corresponding Female Names

Then we mapped each male name to a female name and vice versa. For example, we swapped sentences containing “Micheal” with “Jennifer” along with swapping other gender pronouns as necessary. But since there were not many sentences containing the occupations and these selected names, we had to augment data to increase the dataset. So, after finding a sentence containing a male/female name along with occupation, we applied CDA as before and then we reproduced this sentence 29 more times for all the male and female pairs. Through this method we found 18676 sentences. Among these sentences, 6288 (3144×2) sentences were chosen for training to be consistent with the other two experiments.

An example sentence-pair:

She pointed to her treasury secretary, Cynthia Geithner, and told me, You should give this feminine some tips.

He pointed to his treasury secretary, Timothy Geithner, and told me, You should give this guy some tips.

Then we reproduced this sentence 29 more times containing all the male and female name pairs (like Micheal-Jennifer, David-Linda etc.)

Chapter 4

Methodology

4.1 Model Description

4.1.1 Introduction of BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a pre-trained deep-learning language model that Google AI Language has created. BERT is one of the most widely used and powerful models in the field of natural language processing (NLP) and has achieved groundbreaking results in a wide range of NLP tasks such as question answering, text classification, and language generation.

BERT is based on the Transformer architecture, which is a deep neural network framework that employs self-attention mechanisms to handle data sequences like text. In this model, each input and output element is connected, and the weights between them are dynamically determined based on context and relationship. According to Muller [44], it is pre-trained using text from Wikipedia and Google's BooksCorpus, which is roughly adjusted to 3.3 billion words. Even though training a huge amount of data like this would have taken a large amount of time, but the training time was much lower due to the transformer architecture and using TPUs (tensor processing units). 64 TPUs trained BERT in a matter of 4 days.

One of the key features of BERT is its ability to understand the context of words in a sentence by taking into account the words that come before and after them [44]. Prior to now, language models could only scan text input sequentially, that is, either from right to left or from left to right, but they could not do both at the same time. The primary technological advancement of BERT is its bidirectionality. Hence, BERT can read sentences from both directions. This was made possible with the help of Transformers, a popular attention model. A model's bidirectionality is crucial for fully comprehending a language's meaning. As a result, BERT is able to capture deeper and more detailed comprehensions of how language works.

According to Alammar [10], BERT is a stack of encoders from the Transformer architecture. Each encoder consists of self-attention and feedforward network. Self-attention layer connects a word to all other words in a sentence which helps to capture the context while Feed forward network introduces non-linearity and improves

the model’s ability to grasp complex connections between representations. Before training the data, a certain amount of text preprocessing needs to be done. There are basically three types of information embedded into the input ([44]. At first, we have positional embedding, where the location of words in a phrase is learned and subsequently expressed by BERT. This step is required so that the transformer can successfully record sequence information. Secondly, segment embedding is done so that BERT can develop a distinctive embedding that distinguishes between two sentences side by side. Lastly, token embedding takes place, where words are represented in a numerical way.

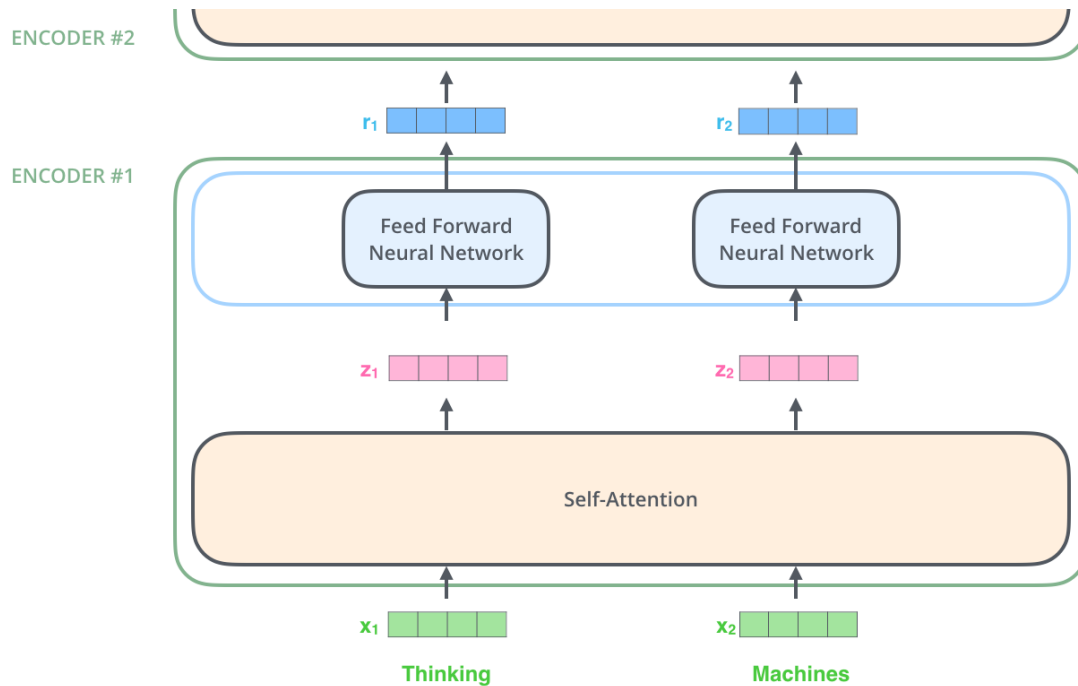


Figure 4.1: BERT architecture [10]

The pre-training of BERT involves two major techniques, which are Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) [23].

In masked language modeling (MLM), the model is trained in such a way that it can predict the words that were masked in a given sentence. In addition to it, the context of the surrounding words should also be taken into consideration. Hence, the model randomly masks some percentage of the input tokens and then attempts to predict the masked tokens. This therefore forces the model to understand the context in which the words are being used so that it can generate contextualized embeddings.

- Select 15% of the tokens.
- Replace 80% of the selected tokens with [MASK].
- Replace 10% of the selected tokens with a random word.
- Keep the rest 10% of the selected tokens as it is.

In next sentence prediction (NSP), the model is trained to predict whether a pair of sentences are consecutive or not. By doing so, the model is able to comprehend how sentences are related to one another and produce embeddings that represent the sense of complete sentences rather than just individual words. In training, BERT’s next sentence prediction accuracy is improved by exposing it to a mixture of 50% right sentence pairs and 50% random sentence pairs.

As soon as the model has been pre-trained on an extensive amount of text data, it can be fine-tuned for a particular downstream job, such as sentiment analysis, question response, or machine translation. The pre-trained model is further fine-tuned by adding a task-specific output layer, and the model is then trained on a smaller collection of annotated data for the particular task.

4.1.2 Variations of BERT

Comparison	BERT	RoBERTa	DistilBERT	ALBERT
Parameters	Base: 110M Large: 340M	Base:125M Large: 355M	Base: 66M	Base:12M Large:18M
Layers/Hidden dimensions/Self-Attention Heads	Base: 12 / 768 / 12 Large: 24 / 1024 / 16	Base: 12 / 768 / 12 Large: 24 / 1024 / 16	Base: 6 / 768 / 12	Base: 12 / 768 / 12 Large: 24 / 1024 / 16
Pre-training Data	BooksCorpus + English Wikipedia = 16GB	BERT + CCNews + OpenWebText + Stories = 160 GB	BooksCorpus + English Wikipedia = 16GB	BooksCorpus + English Wikipedia = 16GB
Method	Bidirectional Transformer, MLM & NSP	BERT without NSP, Using Dynamic Masking	BERT Distillation	BERT with reduced parameters & SOP

Table 4.1: Difference between different BERT variants [41]

BERT has several variations that have been developed to address specific NLP tasks. We have performed our experiments on the following models:

- BERT-Base-Uncased: This is the original BERT model architecture. With 12 transformer layers, 12 attention heads and 110 million parameters, this is the simplest type of BERT. It was trained on an extensive collection of uncased English text, which means that there is no difference between capitalized and non-capitalized words and that all of the text is in lowercase. This model is an excellent starting point for many NLP tasks as it is reasonably quick to train compared to other Bert models and has a fairly high degree of accuracy [23].
- BERT-Large-Uncased: With 24 transformer layers, 16 attention heads and 340 million parameters, this is a larger and more intricate variant of the BERT. It is more computationally expensive to train and use than BERT Base [23].
- RoBERTa-Base: With 12 transformer layers, 12 attention heads and 125 million parameters, it is trained on a combination of cased and uncased English text. This BERT variant was trained using a larger corpus of text and a different pre-training purpose. This model is trained using Dynamic Masking unlike BERT where static masking was used, trained without Next Sentence Prediction (NSP) and trained on large mini-batches [27].
- RoBERTa-Large: It is an extended version of RoBERTa-Base with 24 transformer layers, 16 attention heads and 355 million parameters. It has a greater

ability to recognize more intricate linguistic patterns but is more computationally costly to train and use [27].

- DistilBERT-Base-Uncased: It is a more compact and effective variant of the BERT with only 6 transformer layers, 12 attention heads and 66 million parameters. While distillation is used to compress the knowledge from a larger pre-trained model into a smaller model, it is also trained on uncased English text. Compared to the bigger BERT models, this model is quicker and uses fewer computational resources, but it may not be as accurate in some situations [30].
- ALBERT-Base-v1: It is “A Lite version of BERT” intended to cut down on the amount of parameters and boost training effectiveness without sacrificing performance. It has 12 transformer layers, 12 attention heads and 12 million parameters. It is primarily pretrained on uncased English text data. It utilizes factorized embedding parameterization and cross-layer parameter sharing to overcome huge parameters in BERT [26].
- ALBERT-Large-v1: ALBERT Large is an enlarged version of the original ALBERT that can store more data. It has 18 million adjustable parameters, 24 transformer layers, and 16 monitoring nodes. When compared to ALBERT Base, ALBERT Large offers improved performance but consumes more processing resources [26].

4.2 Cosine Similarity Test

Cosine similarity is used in NLP and information retrieval to compare text documents, which are represented as vectors of word frequencies or embeddings. It is used in machine learning clustering, classification, and recommendation systems. BERT uses “sentence encoding” to represent text in a high-dimensional space as vectors. Texts with similar meanings and situations should have vector representations in this space that are near.

Cosine similarity and Euclidean distance are popular distance metrics in natural language processing and machine learning. In many situations, cosine similarity is superior to Euclidean distance because it better captures vector similarity in high-dimensional domains. Cosine similarity is insensitive to vector magnitude. We have used cosine similarity since it is more effective than Euclidean distance in many cases.

We selected pairs of texts that were similar in meaning but different in gender to perform the cosine similarity test on BERT. We examined the cosine similarity between two occupations used in two sentences that only varied in gender (e.g., “he” vs. “she”). For example, the pair of sentences are “The nurse notified the patient that his shift would be ending in an hour” and “The nurse notified the patient that her shift would be ending in an hour”. The cosine similarity between “nurse” in the first sentence and “nurse” in the second sentence was calculated. These sentences were taken from the Winogender dataset [16]. For each 60 occupations, there are 4 pairs of sentences. For each pair, cosine similarity between the occupation words is calculated and then averaged among the 4 pairs to get a better representation of

the similarity of that specific occupation.
for each occupation:

for each pair:

$$\text{similarity} = \text{cosine}(\text{occupation in sentence 1}, \text{occupation in sentence 2})$$

$$\text{avg_similarity} = \frac{\sum \text{similarity}}{\text{number of pairs}}$$

If the cosine similarity is low, it means that occupation (nurse in the example) is giving different embeddings depending on the gender used in the context. This will show that the model is biased when creating embeddings. The cosine similarity of two vectors is 1 if they are identical and in the same direction. If the cosine similarity is 0, the vectors are orthogonal and have no similarity. The vectors are in opposite directions and utterly distinct if the cosine similarity is -1.

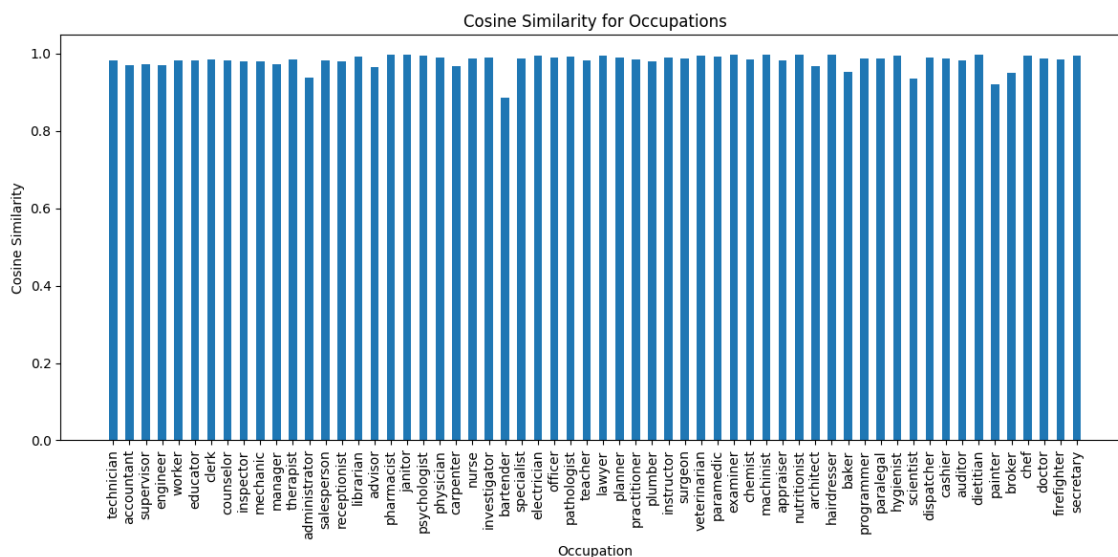


Figure 4.2: Cosine Similarity Test

Our studies showed that most results were between 0.80 and 0.99. This approach did not clearly show gender bias because cosine similarity was high in all cases. Thus, the cosine similarity test did not confirm our discovery. Typical methods that used to work in detecting bias in static word embeddings do not work in contextual word embeddings like BERT which could be due to the ability of the model to capture context and give unique embedding to a word based on the context.

4.3 Direct Bias Test

The direct bias of a set of words is a measure of how close they are to the gender direction vector. On the BERT representations of the professional words in these sentences, we applied the notion of direct bias from Bolukbasi et al. [6]. We used the previously described English-German corpus from WMT18 to compute Direct Bias. With 464,947 lines and 1,004,6125 tokens, we used the English side. We used a set of lists from earlier work to conduct our analysis [6] [24]. The collection of definitional pairings is referred to as the ‘Definitional List’ [6] (for example, she-he,

girl-boy). A ‘Professional List’ [6] is a list of female and male professions (for example, accountant, surgeon).

$$\frac{1}{|N|} \sum_{\omega \in N} \cos(\bar{\omega}, g)$$

Here, N is the number of gender neutral words, g is the gender direction, and \vec{w} is the word vector of each profession

We began by identifying two gendered pronouns or nouns associated with opposing genders, such as “he” and “she” or “man” and “woman”. The ‘Profession List’ is then used to reference a list of male and female professions. To avoid the influence of the gendered nouns over the presence of bias in the former, we removed sentences that had both a professional and a definitional gender word. So, the sentences that only contain profession words from the ‘Profession List’ are then extracted from the big text dataset, WMT18 [17].

The BERT model was then used to generate vector representations or embeddings for each gender-neutral profession and gendered word. Vector representations are numerical representations of word meanings that can be utilised for further investigation. However, the embeddings will be determined by the context in which the words appear. As a result, depending on the context, we will obtain multiple embeddings for the same word.

To create a single vector representation for each word, we averaged the embeddings across all occurrences of each gendered term. In our study, following Bolukbasi et al. [6], we selected “he” and “she” as the gendered term to determine the gender direction. So, we generated embeddings for “he” in all the sentences in the dataset where “he” appeared and averaged it and did the same for “she”. Then we calculate the ‘gender direction’ by subtracting the vector representations of the two gendered terms. In the embedding space, this direction indicates the gender direction. The choice of gendered phrases can influence gender orientation. Using “man” and “woman” as gendered words, for example, may result in a different gender direction than using “he” and “she.”

To calculate the cosine similarity scores between the gender direction and the embeddings of the set of occupational names from ‘Profession List’, such as “doctor,” “teacher”, we just compute the cosine similarity between the gender direction and the embedding of each occupation. Once again, the embeddings of each occupation is taken by averaging the embeddings of the occupation taken from the sentences in the dataset.

Finally, the average cosine similarity score for all gender-neutral professional words is determined which shows the total strength of the relationship in the embedding space between the gender direction and gender-neutral profession nouns.

The mean cosine similarity score is compared to a baseline value. A typical baseline value is the mean cosine similarity score between the gender direction and a random direction in the embedding space. The random direction is generated by averaging 100 random word embeddings created using random function. If mean cosine similarity score is much greater than the baseline value, the model may have gender

bias.

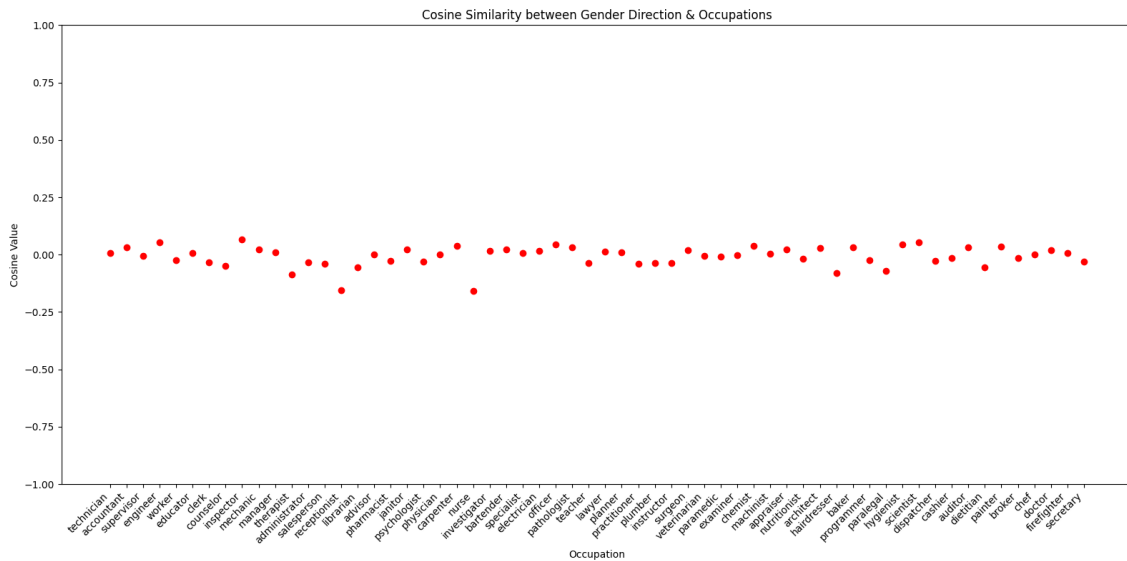


Figure 4.3: Direct Bias Test

Mean cosine similarity score across all gender-neutral words (professions): -0.0073495
Baseline score (mean cosine similarity between gender direction and random directions): -0.0000060131

The mean cosine similarity score across all gender-neutral words (professions) is much less than the baseline score. As a result, we could not conclusively prove that BERT contains gender bias through direct bias test.

4.4 Masking Probability of BERT

BERT model learns contextualized word embeddings using a masked language modeling (MLM) goal. The MLM challenge entails training the model to predict the original masked tokens by randomly masking parts of the tokens in a given input sequence.

BERT completes the MLM process as follows:

To begin processing the data, WordPiece tokenization is used to separate the input string into individual words. The sequence is then altered by inserting a [MASK] token at random locations.

To encode the input sequence, BERT employs a neural network built on transformers. These transformers then provide a series of hidden representations. The transformative neural networks are composed of self-attention and feedforward layers.

BERT is taught during training to determine the unmasked value of masked tokens. The model creates a probability distribution over the whole vocabulary for each masked token and then chooses the token with the highest probability as the anticipated output.

During training, BERT is tweaked so that it predicts token values with the lowest possible cross-entropy loss.

Together, the syntactic and semantic associations between words in a particular phrase are captured in the context-aware word embeddings that BERT learns with the aid of the MLM task.

We used this Masking Language Model (MLM) process by BERT to detect gender bias. In order to determine the probability of the target [MASK], the BERT model calculates a set of numeric values called logits. To convert these logits into a meaningful probability distribution, the softmax function is applied. The softmax function takes the logits as input and produces a probability distribution over the BERT vocabulary specifically for that position in the sentence. This distribution assigns probabilities to each word in the vocabulary, indicating the likelihood of it being the target at that specific position [MASK]. In order to explore the bias properly, we performed two techniques following Kurita et al. [25]. We first mask the gendered term and analyze the probabilities and then we mask the occupation word and analyze the probabilities. We performed these two techniques on the three experiments - “he-she”, “his-her” and “male-female names”

4.4.1 Masking the gendered term

For the first masking technique, we masked the gendered term while keeping the occupation. We used our 51 sentence structures made for “he-she” mentioned before to create a comparison between the probabilities of the gendered terms. For example, For a sentence structure: “[MASK] dreams of being a good [OCC]”. Here, instead of [OCC], we placed occupation words that are gender neutral, taken from the Winogender dataset, which contains 60 occupational words. For demonstration, we kept - ‘engineer’, ‘librarian’, ‘nurse’, ‘surgeon’, ‘programmer’, ‘chef’, ‘scientist’, ‘secretary’, ‘architect’, ‘teacher’. These occupations were selected randomly from the occupation list from Winogender. Now, for a specific occupation, we calculated the probability of the [MASK] being replaced by gendered pronouns - “he” and “she”. For example, the probability of [MASK] being replaced by “he” and “she” in the sentence “[MASK] dreams of being a good engineer”. If the probability of “he” taking place is significantly greater than that of “she”, we can conclude that BERT is filling the [MASK] with gender bias as it is considering a male to be more compatible with the occupation “engineer”. For a stronger claim, we repeated the procedure for 51 structures for each occupation and averaged them.

It should be noted that, even though we applied the aforementioned method to all the models of BERT, it is not accurate for RoBERTa and ALBERT models. It is because the data used to train the models does not have the occupational words that we used, the vocabulary bank does not have all the occupations. Thus when

we give them sentences to predict words in place of [MASK], they are unable to fully understand the context as those sentences contain occupational words that are missing from the models' vocabulary. It can be seen that the models may have replaced the unknown words with some closely related known words available in their vocabulary, however this is not an accurate context and thus it will not give accurate results. It can be seen from the picture below that not all the occupations are available in RoBERTa and ALBERT models.

```

from transformers import AutoModelForMaskedLM, AutoTokenizer

models=["bert-base-uncased", "bert-large-uncased", "distilbert-base-uncased", "roberta-base", "roberta-large", "albert-base-v1", "albert-large-v1"]
occupations=["engineer", "librarian", "nurse", "surgeon", "programmer", "chef", "scientist", "secretary", "architect", "teacher"]

for model_checkpoint in models:
    print(model_checkpoint, ":")
    tokenizer = AutoTokenizer.from_pretrained(model_checkpoint)
    model_vocab = tokenizer.get_vocab()

    missing_occupations = []
    for occupation in occupations:
        if occupation not in model_vocab:
            missing_occupations.append(occupation)

    if len(missing_occupations) == 0:
        print(f"All occupations are available in the vocabulary")
    else:
        print(f"Occupations available in the vocabulary:", end=" ")
        print(' '.join(occupation for occupation in occupations if occupation not in missing_occupations))
        print("Occupations not available:", end=" ")
        print(' '.join(missing_occupation for missing_occupation in missing_occupations))
    print()

bert-base-uncased :
All occupations are available in the vocabulary

bert-large-uncased :
All occupations are available in the vocabulary

distilbert-base-uncased :
All occupations are available in the vocabulary

roberta-base :
Occupations available in the vocabulary:
Occupations not available: engineer, librarian, nurse, surgeon, programmer, chef, scientist, secretary, architect, teacher

roberta-large :
Occupations available in the vocabulary:
Occupations not available: engineer, librarian, nurse, surgeon, programmer, chef, scientist, secretary, architect, teacher

albert-base-v1 :
Occupations available in the vocabulary: engineer, secretary, architect, teacher
Occupations not available: librarian, nurse, surgeon, programmer, chef, scientist

albert-large-v1 :
Occupations available in the vocabulary: engineer, secretary, architect, teacher
Occupations not available: librarian, nurse, surgeon, programmer, chef, scientist

```

Figure 4.4: Missing Occupations in Vocab

Along with this, RoBERTa and ALBERT models also do not contain the male and female names that we have used for our third experiment. Hence, these models are also not applicable for the first masking technique (masking the gendered term) when it comes to the third experiment (“male-female names”).

```

from transformers import AutoModelForMaskedLM, AutoTokenizer

models=['bert-base-uncased','bert-large-uncased','distilbert-base-uncased','roberta-base','roberta-large','albert-base-v1','albert-large-v1']
male_names = ['Michael','David','James','John','Robert','William','Richard','Thomas','Christopher','Joseph','Steven','Paul','Daniel','Andrew','Kenneth','George','Charles','Stephen','Anthony','Edward','Ronald','Kevin','Matthew','Jason','Timothy','Gary','Jeffrey','Scott','Jennifer','Patricia','Susan','Sarah','Jessica']
female_names = ['Jennifer','Linda','Patricia','Susan','Mary','Sarah','Jessica','Elizabeth','Karen','Nancy','Lisa','Margaret','Betty','Sandra','Ashley','Dorothy','Michelle']
male_names = [name.lower() for name in male_names]
female_names = [name.lower() for name in female_names]
all_names = male_names+female_names

for model_checkpoint in models:
    print(model_checkpoint,":")
    tokenizer = AutoTokenizer.from_pretrained(model_checkpoint)
    model_vocab = tokenizer.get_vocab()

    missing_names = []
    for name in all_names:
        if name not in model_vocab:
            missing_names.append(name)

    if len(missing_names) == 0:
        print("All names are available in the vocabulary")
    else:
        print("Names available in the vocabulary:",end=" ")
        print(', '.join(name for name in all_names if name not in missing_names))
        print("Names not available:",end=" ")
        print(', '.join(missing_name for missing_name in missing_names))
        print()

bert-base-uncased :
All names are available in the vocabulary

bert-large-uncased :
All names are available in the vocabulary

distilbert-base-uncased :
All names are available in the vocabulary

roberta-base :
Names available in the vocabulary: john, gary, mary, amy
Names not available: michael, david, james, robert, william, richard, thomas, christopher, joseph, steven, paul, daniel, andrew, kenneth, george, charles, stephen, anthony, edward,
roberta-large :
Names available in the vocabulary: john, gary, mary, amy
Names not available: michael, david, james, robert, william, richard, thomas, christopher, joseph, steven, paul, daniel, andrew, kenneth, george, charles, stephen, anthony, edward,
albert-base-v1 :
Names available in the vocabulary: michael, david, james, john, robert, william, richard, thomas, joseph, paul, daniel, andrew, george, charles, brian, linda, mary, lisa
Names not available: christopher, steven, kenneth, stephen, anthony, edward, ronald, kevin, matthew, jason, timothy, gary, jeffrey, scott, jennifer, patricia, susan, sarah, jessica,
albert-large-v1 :
Names available in the vocabulary: michael, david, james, john, robert, william, richard, thomas, joseph, paul, daniel, andrew, george, charles, brian, linda, mary, lisa
Names not available: christopher, steven, kenneth, stephen, anthony, edward, ronald, kevin, matthew, jason, timothy, gary, jeffrey, scott, jennifer, patricia, susan, sarah, jessica,

```

Figure 4.5: Missing Names in Vocab

4.4.2 Masking the occupation

For the second masking technique, we masked the occupation rather than the gendered term. Here also, we used our 51 sentence structures made for “he-she” mentioned before to create a comparison between the probabilities of the gendered terms. For example, For a sentence structure: “[GENDER] dreams of being a good [MASK]”. Here, instead of [GENDER], we placed “he” and “she”. Now, for each gendered pronoun, we calculated the probability of the [MASK] being replaced by the occupations that we selected previously. For example, bias can be identified if the given sentence is “he dreams of being a good [MASK]” and the [MASK] is replaced by engineer by the model by a greater probability than for nurse. For a stronger claim, we repeated the procedure for 51 structures for each occupation and averaged them.

As mentioned before, the vocabulary of RoBERTa and ALBERT models do not contain the occupational words that we are using, so it is unable to replace [MASK] with any of the occupations, thus it is not possible to apply the aforementioned technique in the RoBERTa and ALBERT models.

4.5 Visualization of Masking Probability

4.5.1 Masking the gendered term for exp 1: he-she

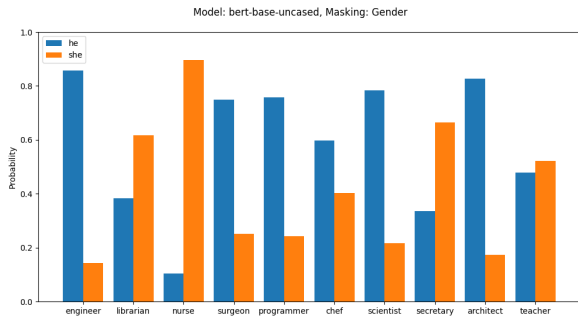


Figure 4.6: BERT Base Mask Gender exp 1

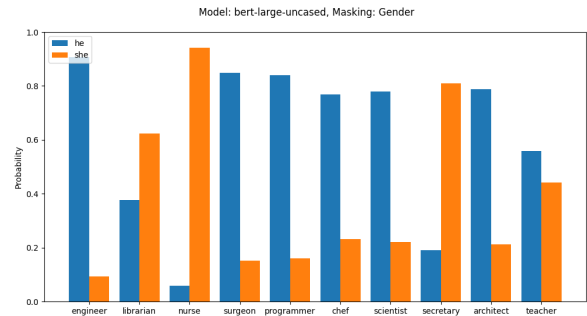


Figure 4.7: BERT Large Mask Gender exp 1

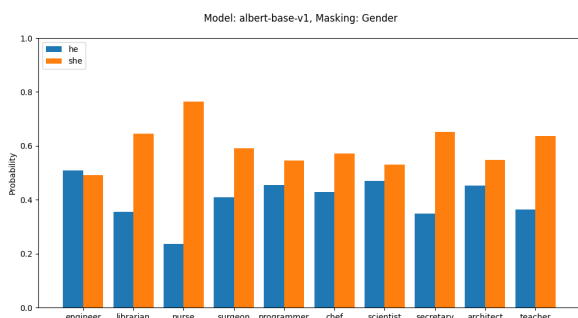


Figure 4.8: ALBERT Base Mask Gender exp 1

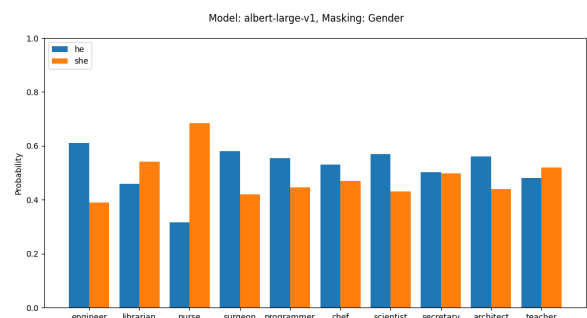


Figure 4.9: ALBERT Large Mask Gender exp 1

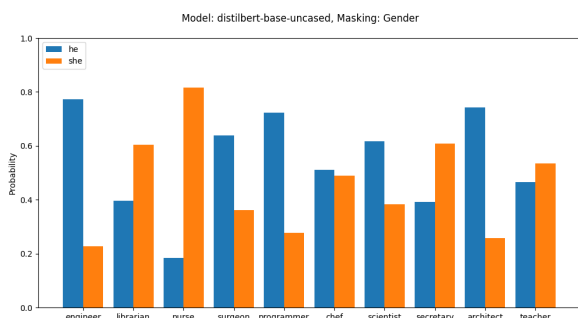


Figure 4.10: DistilBERT Base Mask Gender exp 1

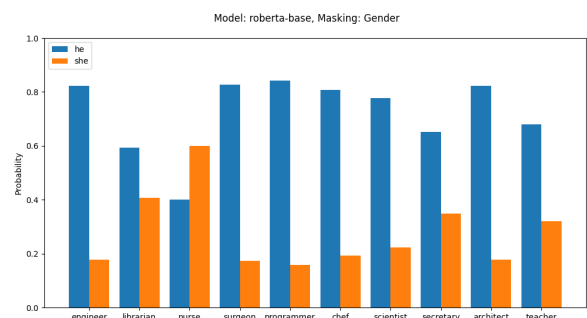


Figure 4.11: RoBERTa Base Mask Gender exp 1

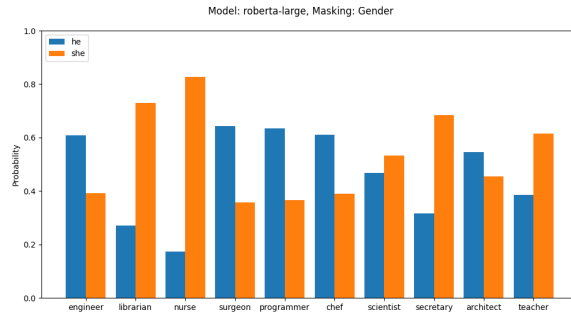


Figure 4.12: RoBERTa Large Mask Gender exp 1

4.5.2 Masking the occupation for exp 1: he-she

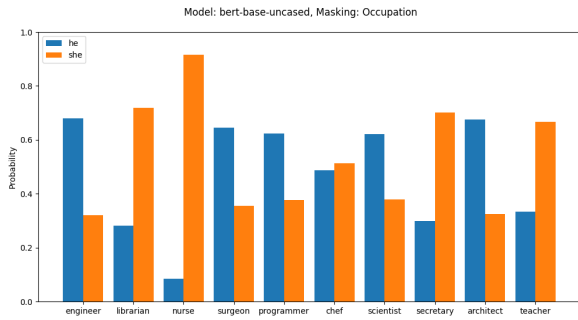


Figure 4.13: BERT Base Mask Occ exp 1

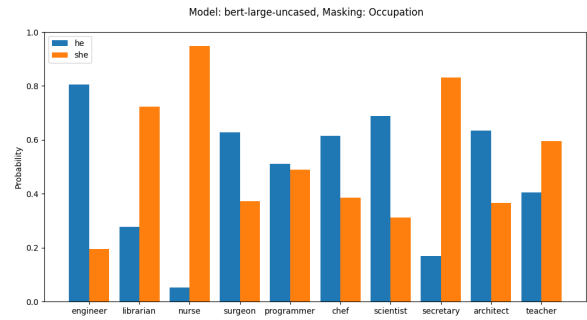


Figure 4.14: BERT Large Mask Occ exp 1

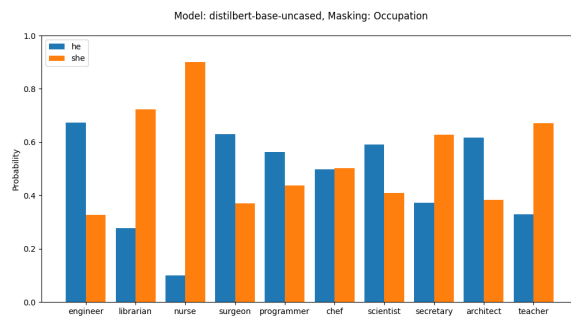


Figure 4.15: DistilBERT Base Mask Occ exp 1

4.5.3 Masking the gendered term for exp 2: his-her

The same procedure is repeated for this experiment. The 51 sentence structure for “his-her” is used and probability of [MASK] being replaced by “his” and “her” is calculated and averaged over all structures for each occupation.

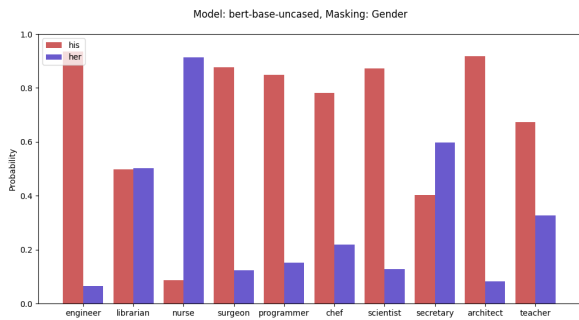


Figure 4.16: BERT Base Mask Gender exp 2

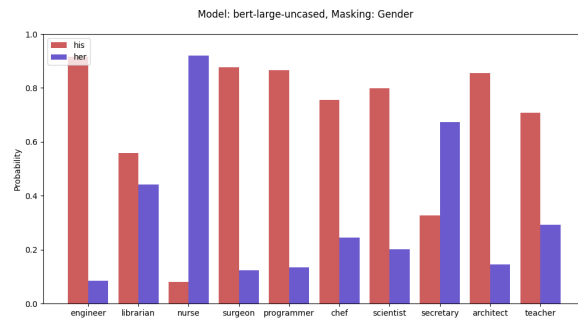


Figure 4.17: BERT Large Mask Gender exp 2

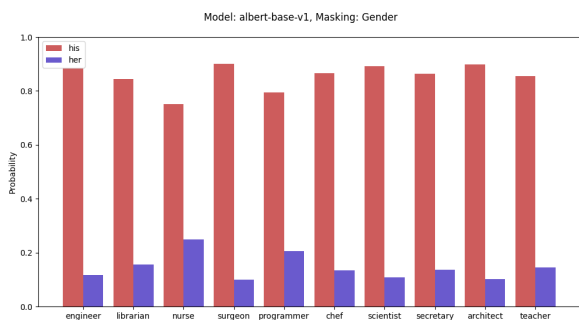


Figure 4.18: ALBERT Base Mask Gender exp 2

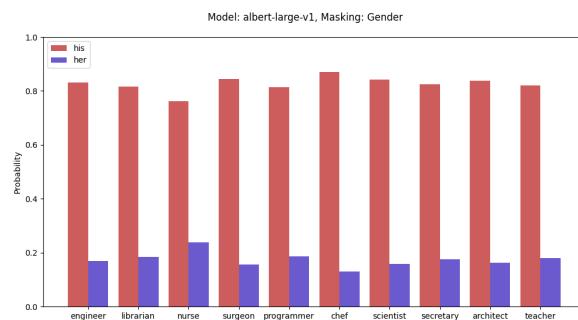


Figure 4.19: ALBERT Large Mask Occ exp 2

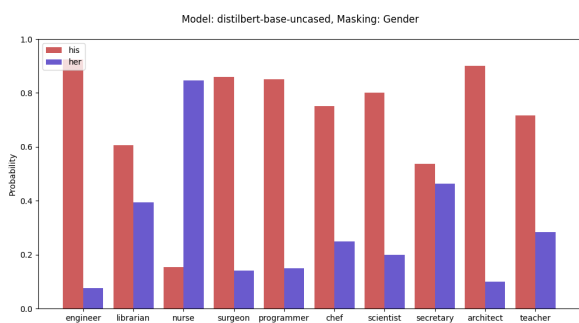


Figure 4.20: DistilBERT Base Mask Gender exp 2

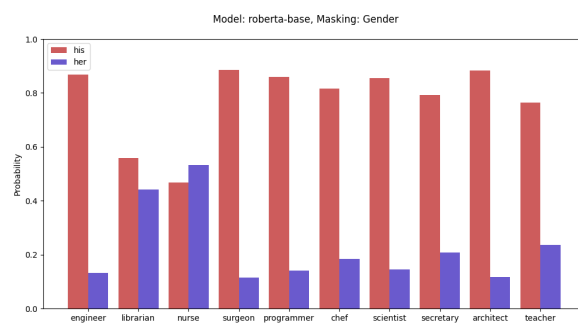


Figure 4.21: RoBERTa Base Mask Gender exp 2

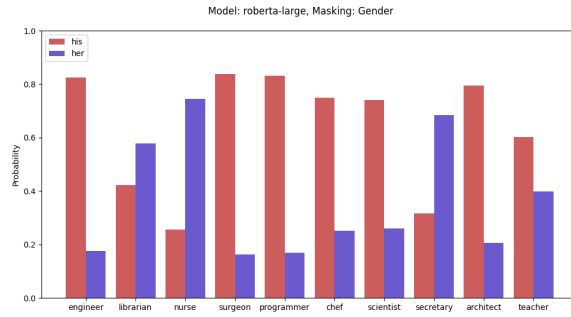


Figure 4.22: RoBERTa Large Mask Gender exp 2

4.5.4 Masking the occupation for exp 2: his-her

The same procedure is repeated for this experiment. The 51 sentence structures for “his-her” is used and probability of [MASK] being replaced by the occupations is calculated and averaged over all structures for each occupation. This is not applicable for RoBERTa and ALBERT as they do not contain the occupations in their vocabulary which is shown before.

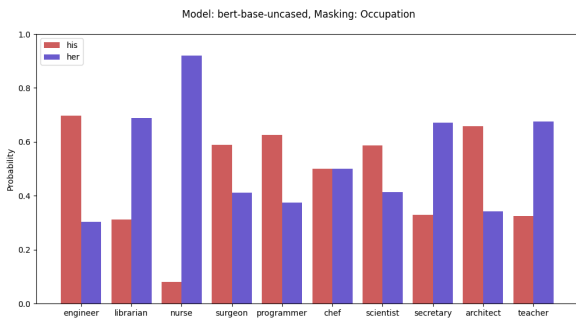


Figure 4.23: BERT Base Mask Occ exp 2

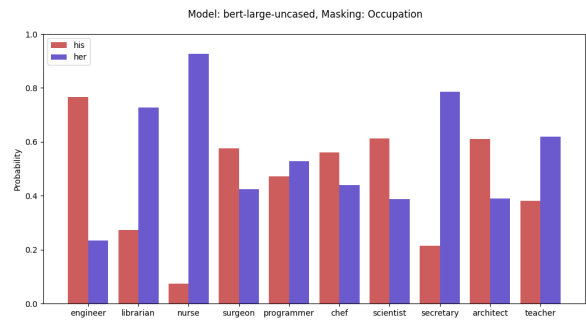


Figure 4.24: BERT Base Mask Occ exp 2

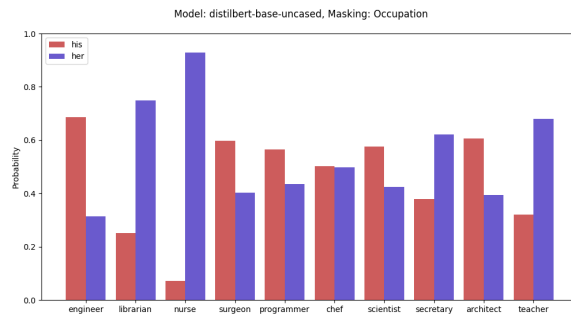


Figure 4.25: DistilBERT Base Mask Occ exp 2

4.5.5 Masking the gendered term for exp 3: male-female name

Here, since we worked with 29 male and female names as mentioned earlier, we calculated the probability of the [MASK] being replaced by 29 male names and averaged them to get an overall probability for [MASK] being replaced by a male name and did the same process for the female names. Like before, we then averaged over all the sentence structures for each occupation. This is not applicable for RoBERTa and ALBERT as they do not contain these names in their vocabulary which is shown before.

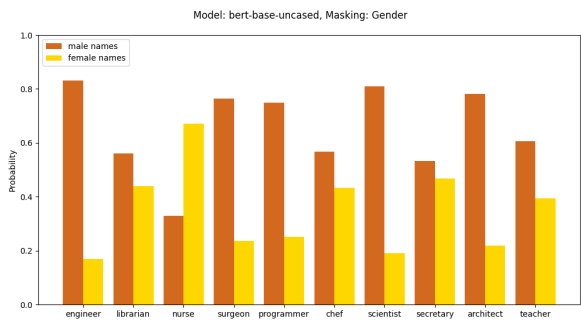


Figure 4.26: BERT Base Mask Gender exp 3

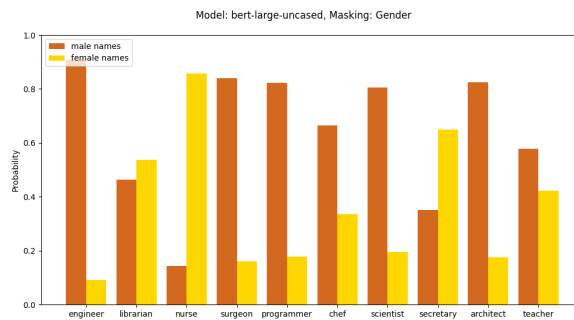


Figure 4.27: BERT Large Mask Gender exp 3

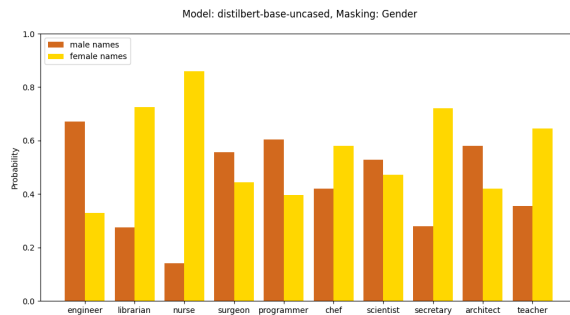


Figure 4.28: DistilBERT Base Mask Gender exp 3

4.5.6 Masking the occupation for exp 3: male-female name

Here, since we worked with 29 male and female names, we calculated the probability of the [MASK] being replaced by an occupation for 29 male names for a fixed sentence structure and averaged them to get an overall probability for [MASK] being replaced by an occupation for male name and did the same process for the female names. Next, we averaged over all the sentence structures for each occupation to get a better representation. This is not applicable for RoBERTa and ALBERT as they do not contain the occupations in their vocabulary which is shown before.

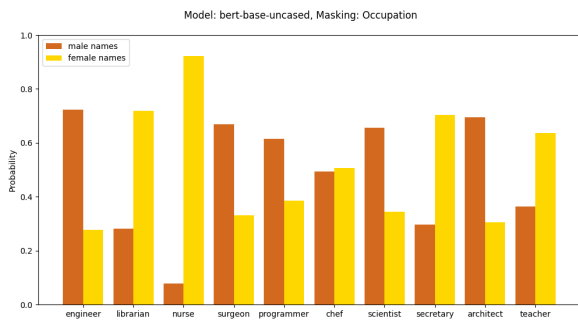


Figure 4.29: BERT Base Mask Occ exp 3

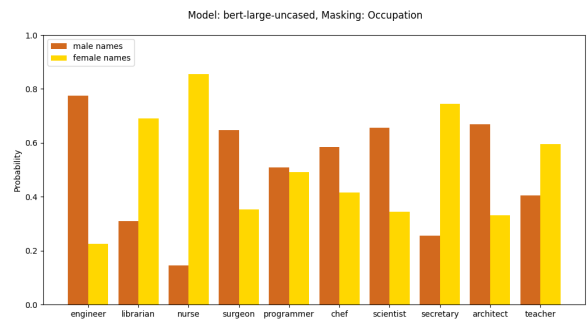


Figure 4.30: BERT Large Mask Gender exp 3

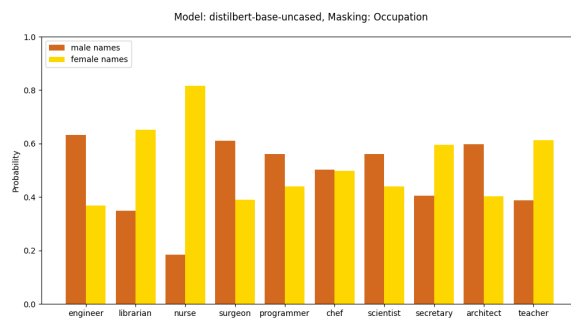


Figure 4.31: DistilBERT Base Mask Gender exp 3

Chapter 5

Debiasing Technique

5.1 Bias Evaluation Metric - MALoR

For selecting a suitable measurement tool through which we can understand how the models are being debiased, we introduced a metric that reliably compares gendered words with occupations and identifies instances of gender bias. We named our metric “MALoR” - Mean Absolute Log of Ratio. Our metric is created on the basis of a variety of different sentence structures used and a list of occupations to test the bias against. For our work, we used the previously mentioned 51 structures and 60 occupations for the metric but it can be modified to have more or different sentence structures and occupations that one may require for a task. Our metric can be applied to any transformer models that support Masked Language Modeling (MLM) hence it can be applied to all the models we are working on.

$$\frac{1}{M} \sum_{j \in \text{occ}} \left| \left(\frac{1}{N} \sum_{i \in \text{sent}} \log_2 \frac{P(\text{male_term})_{ij}}{P(\text{female_term})_{ij}} \right) \right|$$

Here, sent is the sentence structures used, occ is occupational words used. N is the number of sentence structures. M is the number of occupational words used.

For each occupation, the probability of the [MASK] being replaced by a male term and probability of the [MASK] being replaced by a female term are calculated. Then their ratio is calculated to see how greater the probability of the male term is than that of the female term. We applied log base 2 to this ratio to normalize the value and also with using log, if the value is 0, we can say the probability of male term and female term is equal and hence no bias exists. Base 2 of log is used as we are dealing with 2 terms. For better accuracy, we averaged this log ratio over all the sentence structures. Since, male term is in the numerator and female term is in the denominator, if probability of female term is bigger, the log ratio will be negative. Negative log ratio means bias is leaned towards female and positive means bias is leaned towards male. Then we applied absolute mean of this averaged log ratio over all the occupations used which gives us a single value. Mean is used for overall representation for all occupations and Absolute is used because for some occupations, the average log ratio can be positive and for others, it can be negative so taking mean without absolute can neutralize it and give us a wrong representation of bias. With taking absolute, all the averaged log ratio for all the

sentence structures and for all the occupations will be positive and this single value can range from 0 to infinity. 0 indicates no bias and infinite indicates infinite bias. Our aim is to reduce the bias to close to 0. For the three experiments - “he-she”, “his-her” and “male names-female names”, we adjusted our metric accordingly and they are shown below.

5.1.1 Metric for gendered term - he and she

We calculated the absolute mean of averaged log ratio of probability of [MASK] being replaced by “he” to probability of [MASK] being replaced by “she” for all sentence structures and for all occupations. Here the sentences structures created for “he-she” are used.

$$\frac{1}{M} \sum_{j \in \text{occ}} \left| \left(\frac{1}{N} \sum_{i \in \text{sent}} \log_2 \frac{P(\text{he})_{ij}}{P(\text{she})_{ij}} \right) \right|$$

sent is the sentence structures used, occ is occupational words used. N is the number of sentence structures. M is the number of occupational words used.

5.1.2 Metric for gendered term - his and her

Similarly, we calculated the absolute mean of averaged log ratio of probability of [MASK] being replaced by “his” to probability of [MASK] being replaced by “her” for all sentence structures and for all occupations. Here the sentence structures created for “his-her” are used.

$$\frac{1}{M} \sum_{j \in \text{occ}} \left| \left(\frac{1}{N} \sum_{i \in \text{sent}} \log_2 \frac{P(\text{his})_{ij}}{P(\text{her})_{ij}} \right) \right|$$

sent is the sentence structures used, occ is occupational words used. N is the number of sentence structures. M is the number of occupational words used.

5.1.3 Metric for gendered term - male names and female names

For male names and female names, the process is slightly different as we are not dealing with 2 gender terms. We considered 29 male names and 29 female names to calculate the score. To calculate the log ratio, probability of the male names is calculated by averaging all the probabilities of [MASK] being replaced by each male name used and probability of the female names is calculated by averaging all the probabilities of [MASK] being replaced by each female name used. The rest of the process is similar to the previous metrics, we calculated the absolute mean of averaged log ratio of probability of [MASK] being replaced by male names to probability of [MASK] being replaced by female names for all sentence structures and for all occupations. Here the sentence structures created for “male names-female names” are used.

$$\frac{1}{M} \sum_{j \in \text{occ}} \left| \left(\frac{1}{N} \sum_{i \in \text{sent}} \log_2 \frac{P(\text{male_names})_{ij}}{P(\text{female_names})_{ij}} \right) \right|$$

sent is the sentence structures used, occ is occupational words used. N is the number of sentence structures. M is the number of occupational words used.

$$P(\text{male_names}) = \frac{1}{n} \sum_{i \in \text{male}} P(i)$$

male is the most common male names used. n is the number of male names.

$$P(\text{female_names}) = \frac{1}{n} \sum_{i \in \text{female}} P(i)$$

female is the most common female names used. n is the number of female names.

MALoR Score is calculated using all the models and using the 3 experiments. RoBERTa and ALBERT does not support this metric when it comes to male-female names as their vocabulary does not contain the most common male and female names we chose. It can be seen that in all models, there is significant bias as the MALoR scores are not very close to zero.

Model	he-she	his-her	male-female name
bert-base-uncased	1.27	2.51	1.37
bert-large-uncased	1.98	2.55	1.82
distilbert-base-uncased	0.632	2.087	0.604
roberta-base	1.642	1.581	N/A
roberta-large	0.789	1.811	N/A
albert-base-v1	0.619	2.583	N/A
albert-large-v1	0.250	2.255	N/A

Table 5.1: MALoR Scores of different models

5.2 Continue Pre training for Debiasing

5.2.1 Preprocessing

For debiasing the models, we have continued the pretraining process through which BERT was initially trained. We continued the pretraining on the gender balanced dataset mentioned before so that the model learns that the occupation terms can be equally associated with both the gender.

At first, the checkpoint of the model and tokenizer is loaded using AutoModelForMaskedLM and AutoTokenizer, respectively from the transformers library. Then, we imported our gender balanced dataset and separated the sentences using nltk library and put the sentences into a list.

The length of the input sequence is established initially. The input sequences must be the same length since BERT first processes the inputs as two-dimensional tensors. This is accomplished by either trimming excessively lengthy sequences or

adding [PAD] tokens to excessively short sequences known as padding. However, establishing this fixed sequence length is a prerequisite. It can be achieved by setting `fix_len` to be the same as `max_len`, where `max_len` is the length of the longest input or sentence in the dataset. Following Bartl et al. [35], it is preferable if the fixed sequence length is a power of two, or a number in the form, of 2^n . In particular, we are looking for the least power of two that is higher than or equal to the longest possible sequence:

$$\begin{aligned} fix_len &= 2^n : 2^n \geq max_len \\ n &= \lceil \log_2(max_len) \rceil \end{aligned}$$

Here, `n` is settled upon as the upper bound of the binary logarithm of the longest possible sequence.

First, the string comprising a sentence is broken up by whitespaces to approximate the length of the sequence. Then, the `max_len` is the length of the longest sequence among all inputs. When determining the length of a fixed sequence, it is convenient to use a value that is a power of two to make computations simpler [35].

Next, each sentence is tokenized to corresponding indices in BERT vocabulary using the model’s tokenizer which adds special tokens [CLS] at the beginning of the sentence and [SEP] at the end. Then sentences are then padded to the `fix_len` determined earlier.

Finally, in order to train the model to recognise meaningful tokens and ignore padding ones, the padded and encoded inputs are used to generate attention masks. The dimensions of the input tensors are preserved in the attention mask tensors. Tokens that do not belong to pads are labeled as “1” in the attention mask tensor, whereas pad tokens are labeled as “0” for each index of the input tensor. The model’s self-attention mechanism may be directed towards certain tokens or positions in the input sequence using the attention mask.

5.2.2 Training

The tokenized sentences and attention masks are then randomly batched in sets of 32 for further pretraining. It is necessary to mask the inputs before using them in training with BERT’s MLM. We used the conventional approach for masking inputs, as described by Devlin et al. [23]: randomly selecting 15% of the input tokens, masking 80% of them, replacing 10% of them with an arbitrary word, and leaving the remaining 10% unchanged [37]. “`mask_tokens`” is a preexisting function in code written by Gururangan et al. [37] that does the masking. Inputs are cloned into labels. Then the inputs are modified by masking and replacing with random words following the MLM procedure. The modified inputs and their corresponding labels are used to compute loss. The model gives an output token to a masked token based on the highest probability. The loss is generated using cross-entropy loss between the predicted output token for the masked token and the actual token. This loss is then backpropagated and the weights are adjusted accordingly.

The model is trained for 200 epochs utilizing an AdamW optimizer with a learning

rate of 2×10^{-5} and a linear scheduler. 200 epochs are selected by experimentation with different epochs and finding low convergence rate ($<10\%$) in most models. Learning rate of 2×10^{-5} is chosen as initial as it is a common starting point when finetuning BERT according to studies. Batch size of 32 is taken according to dataset size and availability of memory.

Learning rate determines how much the model’s parameters are to be adjusted. Deep learning models are trained using the optimisation method Adam (Adaptive Moment Estimation) to adjust the model’s parameters. AdamW is an adaptation of the Adam optimizer that applies weight decay to the model’s “weight” parameters and provides better control over the regularization strength. On the other hand, linear scheduler creates a schedule with a learning rate that decreases linearly from the initial learning rate set in the optimizer to 0 [40]. When the weights are far from ideal in the beginning, this method can help the model converge more quickly. As training goes on, the weights are steadily improved.

The learning graph is generated using our evaluation metric MALoR. The learning graph contains x-axis as epochs and y-axis as the MALoR. This helps us to determine when the learning procedure converges.

The same procedure is followed for “he-she”, “his-her” and “male-female name” with the respective dataset.

We could not perform our debiasing experimentation on RoBERTa and ALBERT models because these models have different vocabulary or different tokenization procedure. Due to this, these models do not contain the occupations and male-female names that we have used. Hence, it is not possible to perform debiasing using the third experiment “male-female name” with these models. Furthermore, for the other two experiments - “he-she” and “his-her”, the debiasing experiment would not be accurate since the datasets used for debiasing contain occupation words which would be tokenized or changed into a different token by these two models. As a result, the desired effect of occupations equally distributed by both gender would not be there. Due to lack of time and hardware, we could not look into how to avoid this problem but we will be working on this in the future so that we can debias all the models successfully.

Chapter 6

Results and Discussion

6.1 Learning Graph- Epoch vs MALoR

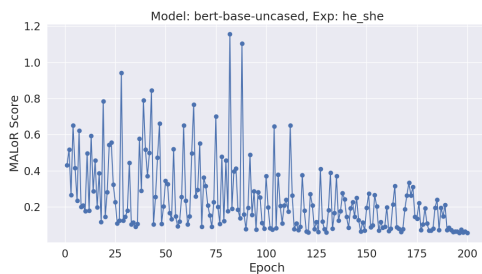


Figure 6.1: BERT Base Graph exp 1

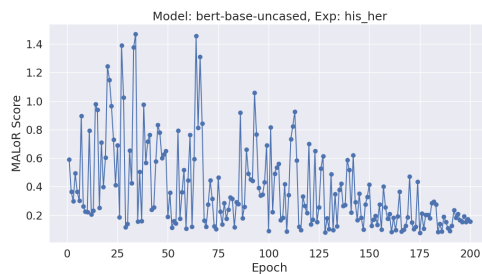


Figure 6.2: BERT Base Graph exp 2

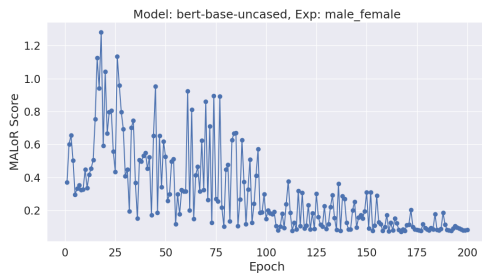


Figure 6.3: BERT Base Graph exp 3

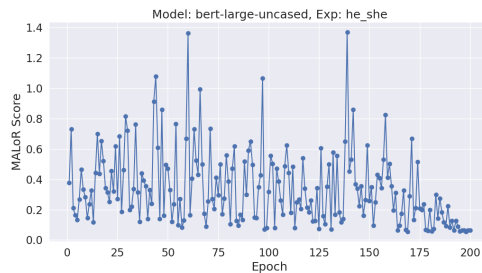


Figure 6.4: BERT Large Graph exp 1

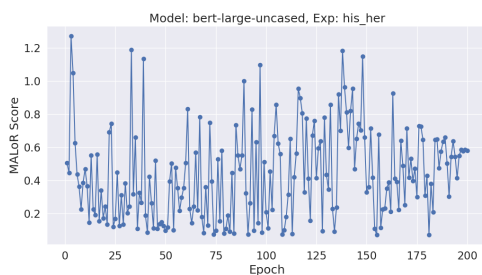


Figure 6.5: BERT Large Graph exp 2

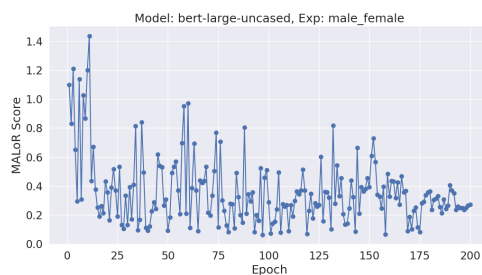


Figure 6.6: BERT Large Graph exp 3

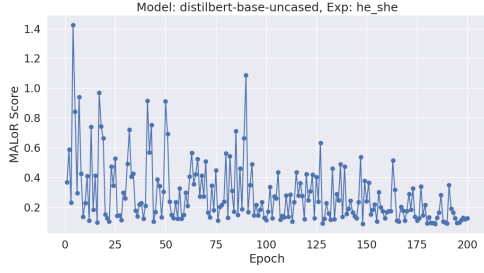


Figure 6.7: DistilBERT Base Graph exp 1

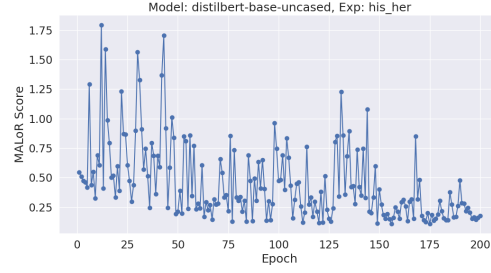


Figure 6.8: DistilBERT Base Graph exp 2

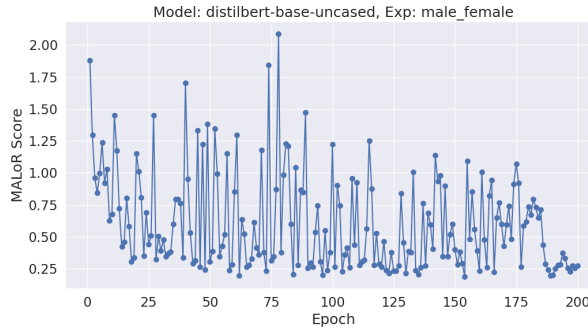


Figure 6.9: DistilBERT Base Graph exp 3

Experiment	Model	Max	Min	Convergence
he-she	bert-base-uncased	0.0843	0.0791	6.62%
	distilbert-base-uncased	0.1451	0.1137	27.55%
	bert-large-uncased	0.0654	0.0634	3.23%
his-her	bert-base-uncased	0.1723	0.1578	9.13%
	distilbert-base-uncased	0.1451	0.1450	22.81%
	bert-large-uncased	0.7777	0.5477	7.01%
male-female name	bert-base-uncased	0.0835	0.0789	5.78%
	distilbert-base-uncased	0.2735	0.2305	18.66%
	bert-large-uncased	0.2715	0.2355	15.29%

Table 6.1: Convergence Rates for last 5 epochs

During training, the model’s performance may see spikes that correspond to brief periods of improvement or decline; however, these fluctuations tend to smooth out as the model converges on a more stable solution with the aid of the scheduler. BERT Base and BERT large models have good convergence as they converge with less than 10%. The convergence rate for male-female name experiment by BERT large was slightly greater. DistilBERT has bad convergence as it does not converge within 10% for any of the 3 experiments. We tried training with different number of epochs but still could not find good convergence. We will work on the convergence in the near future.

6.2 Results of Bias Evaluation Metric MALoR of models before and after debiasing

Model	he-she (before)	he-she (after)
bert-base-uncased	1.275	0.0803 \pm 0.0147
bert-large-uncased	1.979	0.059
distilbert-base-uncased	0.632	0.126 \pm 0.0606

Table 6.2: MALoR Scores of he-she

Model	his-her (before)	his-her (after)
bert-base-uncased	2.514	0.488 \pm 0.224
bert-large-uncased	2.552	0.610
distilbert-base-uncased	2.087	0.179 \pm 0.0684

Table 6.3: MALoR Scores of his-her

Model	male-female (before)	male-female (after)
bert-base-uncased	1.367	0.418 \pm 0.242
bert-large-uncased	1.823	0.338
distilbert-base-uncased	0.604	0.416 \pm 0.157

Table 6.4: MALoR Scores of male-female

The following models were debiased 5 times with different seeds for each experiments and mean MALoR Scores with standard deviation were recorded. Since, there are some randomness involved in the training process, the MALoR Scores have some variance. We could not debias BERT Large multiple times as BERT Large requires much greater GPU RAM and training time which we did not had. Hence, we could only train BERT Large one time for each experiment and could not provide mean and standard deviation for the MALoR Score. We will compute the mean and standard deviation for BERT Large as we get access to higher GPU.

Gender bias has significantly reduced as the initial MALoR Score reduced to a much lower value in all 3 experiments performed by the 3 models.

6.3 Visualization of Masking probability before and after de-biasing

6.3.1 Masking the gendered term for exp 1: he-she

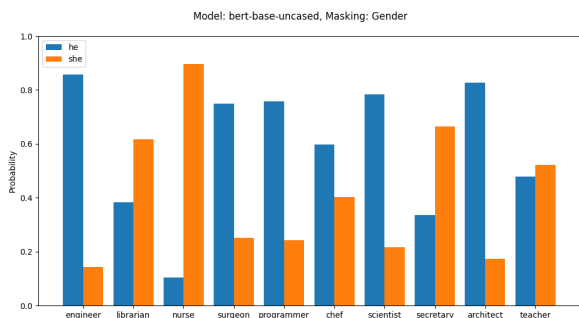


Figure 6.10: Original BERT Base Mask Gender exp 1

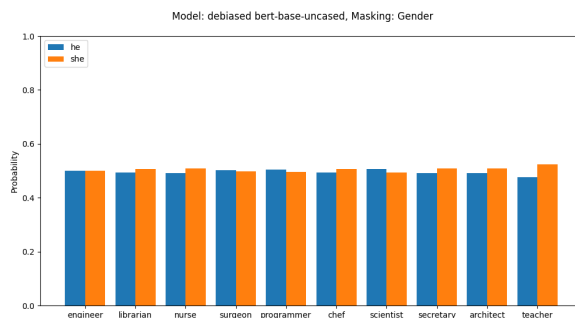


Figure 6.11: Debiased BERT Base Mask Gender exp 1

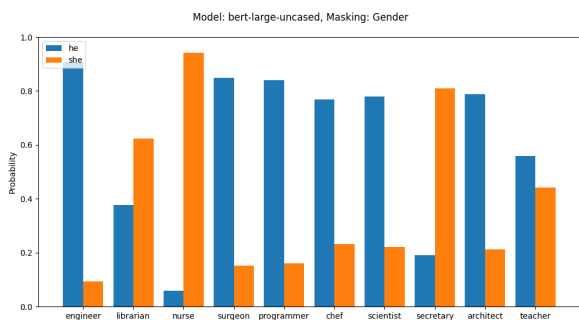


Figure 6.12: Original BERT Large Mask Gender exp 1

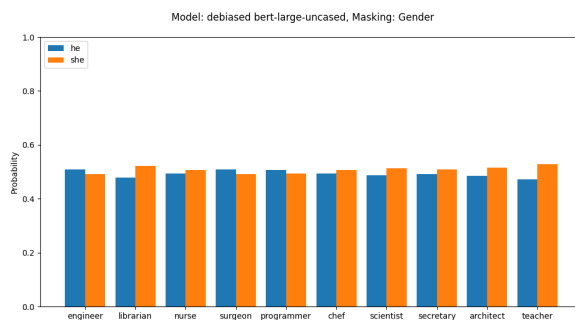


Figure 6.13: Debiased BERT Large Mask Gender exp 1

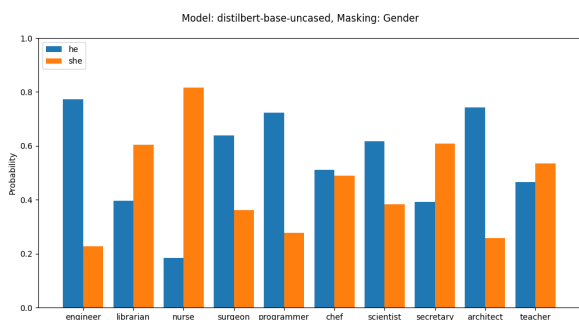


Figure 6.14: Original DistilBERT Mask Gender exp 1

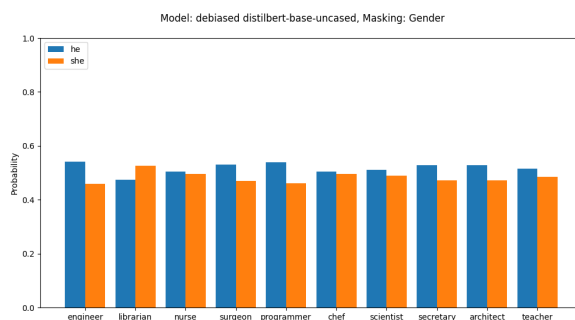


Figure 6.15: Debiased DistilBERT Mask Gender exp 1

6.3.2 Masking the occupation for exp 1: he-she

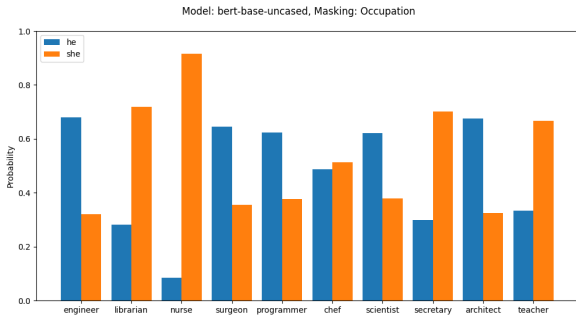


Figure 6.16: Original BERT Base Mask Occ exp 1

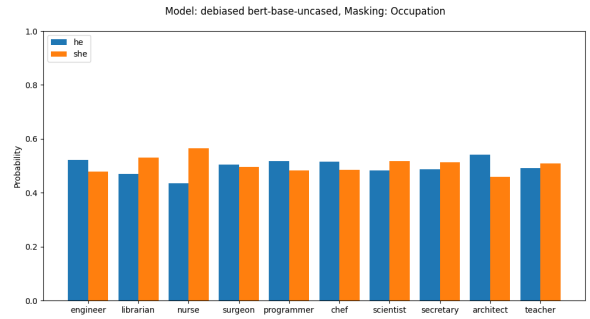


Figure 6.17: Debiased BERT Base Mask Occ exp 1

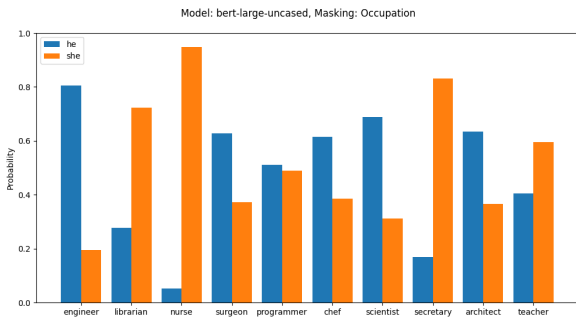


Figure 6.18: Original BERT Large Mask Occ exp 1

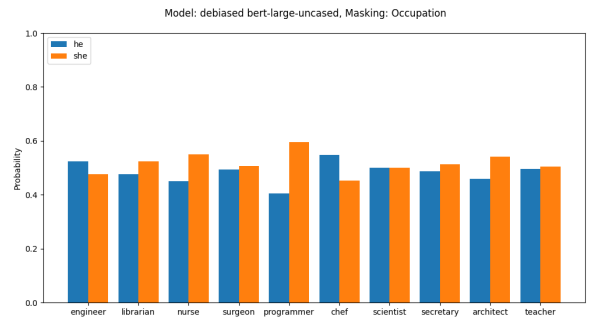


Figure 6.19: Debiased BERT Large Mask Occ exp 1

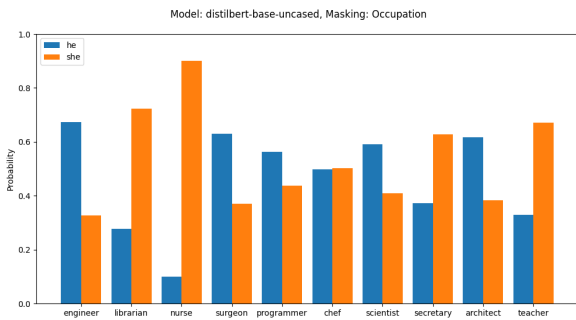


Figure 6.20: Original DistilBERT Mask Occ exp 1

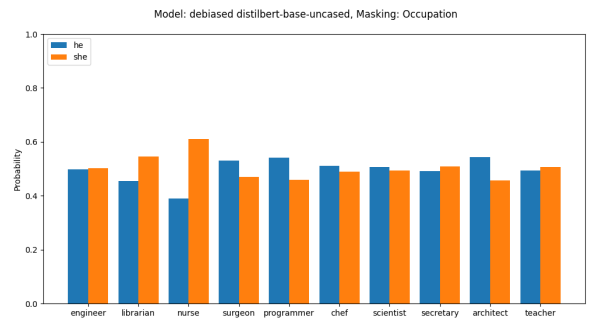


Figure 6.21: Debiased DistilBERT Mask Occ exp 1

6.3.3 Masking the gendered term for exp 2: his-her

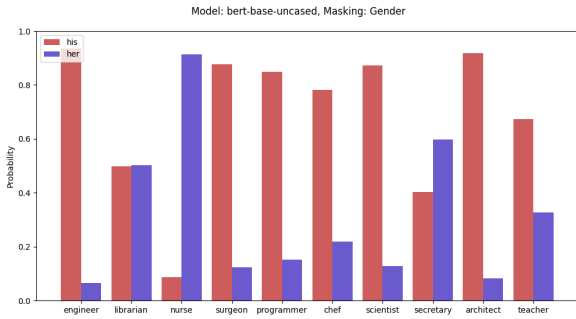


Figure 6.22: Original BERT Base Mask Gender exp 2

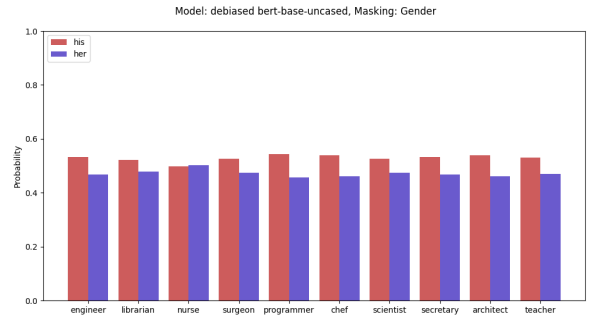


Figure 6.23: Debiased BERT Base Mask Gender exp 2

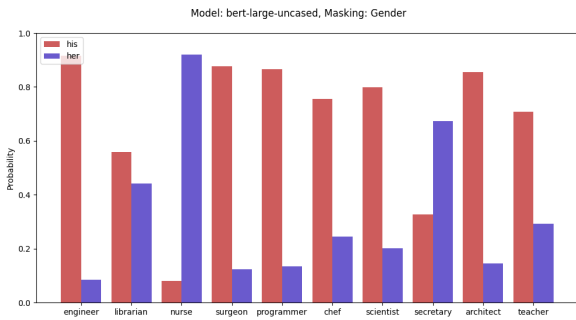


Figure 6.24: Original BERT Large Mask Gender exp 2

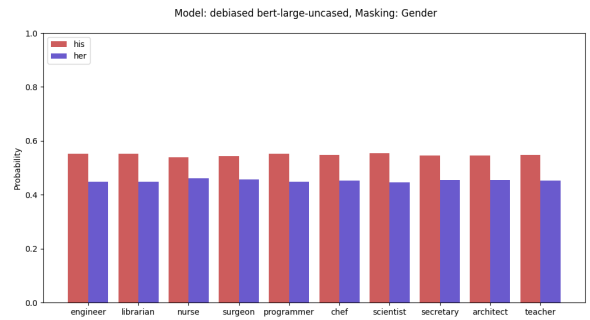


Figure 6.25: Debiased BERT Large Mask Gender exp 2

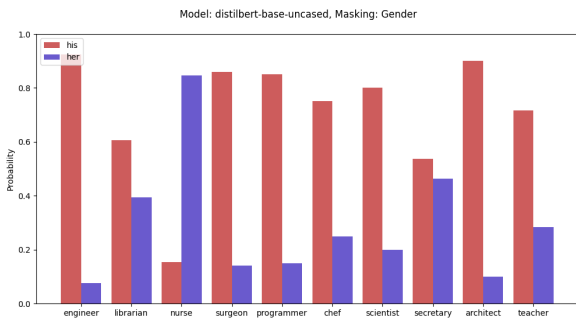


Figure 6.26: Original DistilBERT Mask Gender exp 2

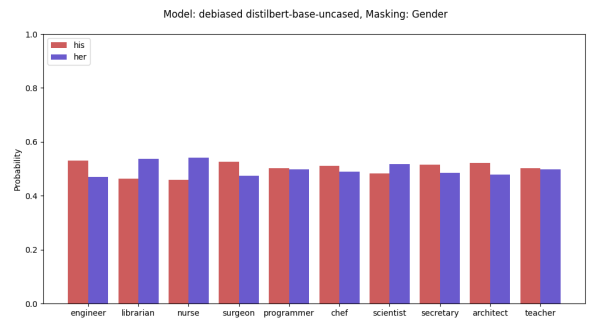


Figure 6.27: Debiased DistilBERT Mask Gender exp 2

6.3.4 Masking the occupation for exp 2: his-her

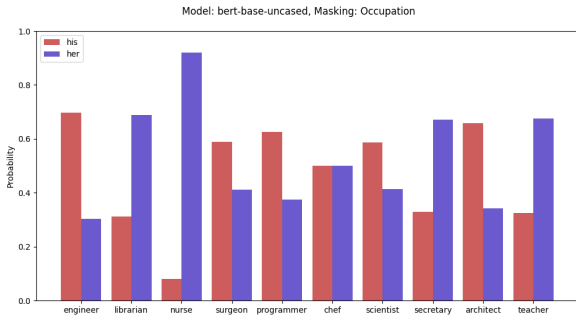


Figure 6.28: Original BERT Base Mask Occ exp 2

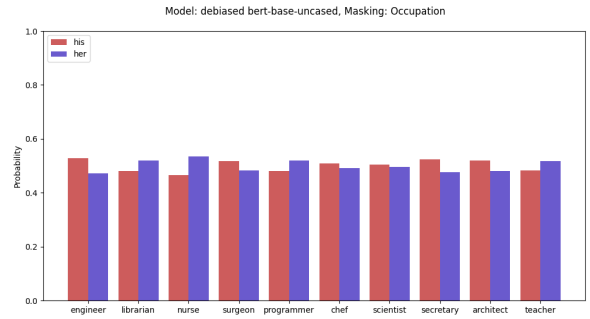


Figure 6.29: Debiased BERT Base Mask Occ exp 2

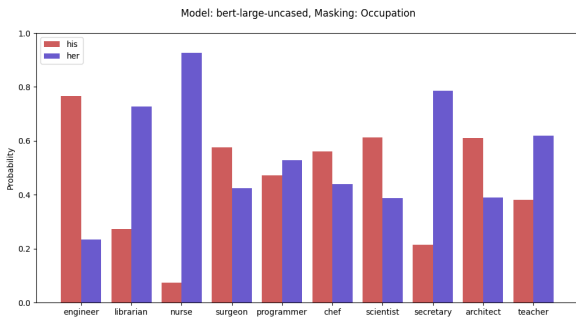


Figure 6.30: Original BERT Large Mask Occ exp 2

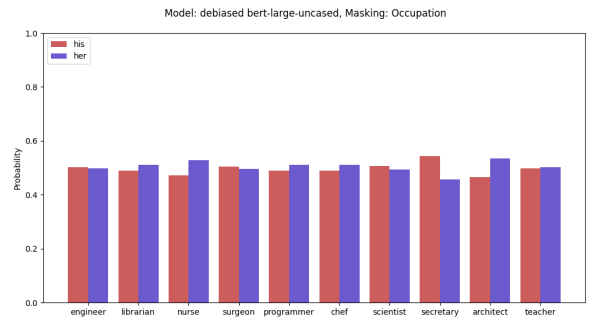


Figure 6.31: Debiased BERT Large Mask Occ exp 2

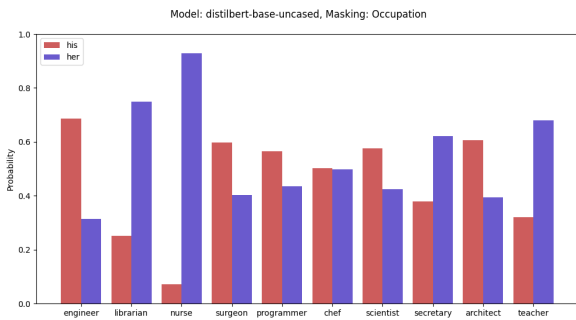


Figure 6.32: Original DistilBERT Mask Occ exp 2

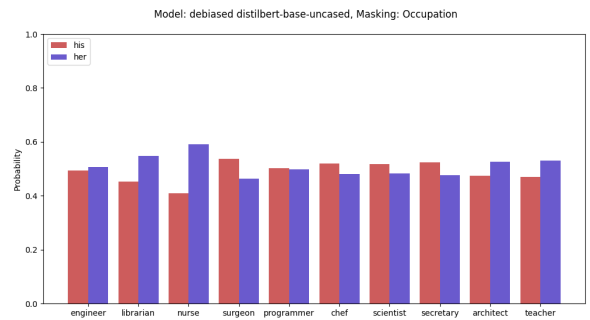


Figure 6.33: Debiased DistilBERT Mask Occ exp 2

6.3.5 Masking the gendered pronoun for exp 3: male-female name

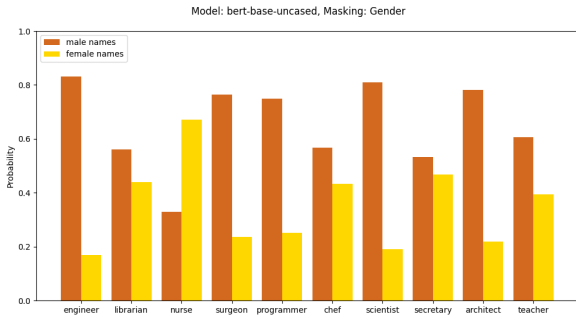


Figure 6.34: Original BERT Base Mask Gender exp 3

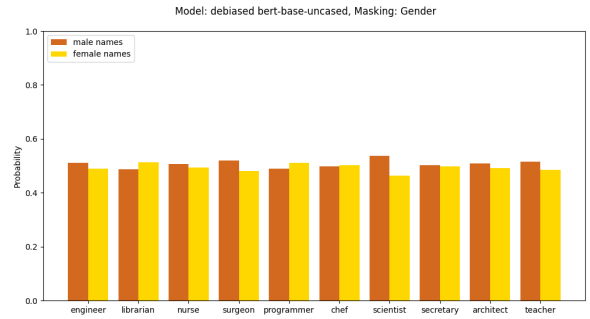


Figure 6.35: Debiased BERT Base Mask Gender exp 3

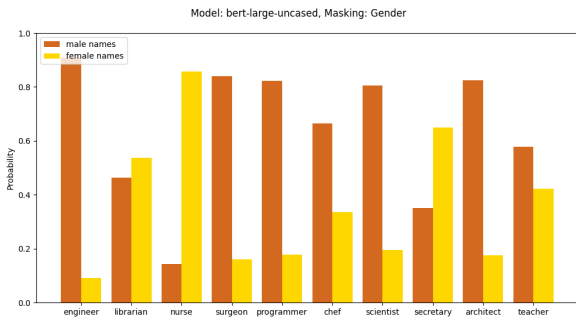


Figure 6.36: Original BERT Large Mask Gender exp 3

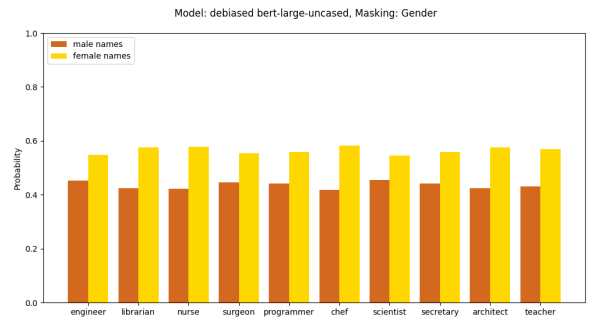


Figure 6.37: Debiased BERT Large Mask Gender exp 3

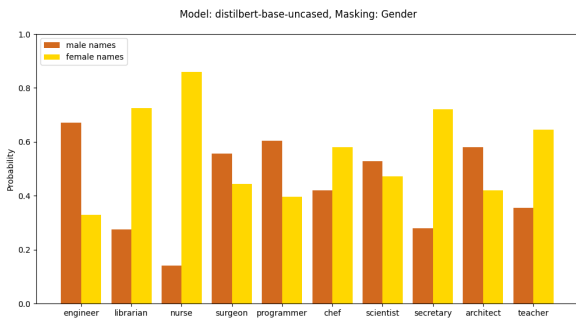


Figure 6.38: Original DistilBERT Mask Gender exp 3

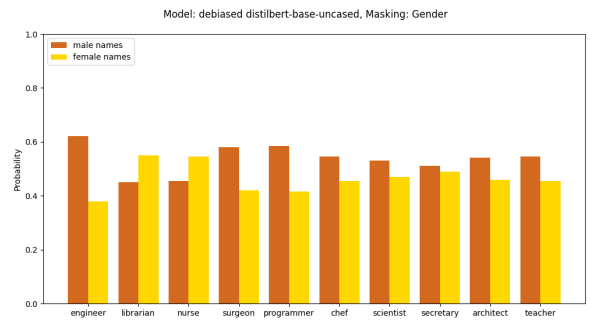


Figure 6.39: Debiased DistilBERT Mask Gender exp 3

6.3.6 Masking the occupation for exp 3: male-female name

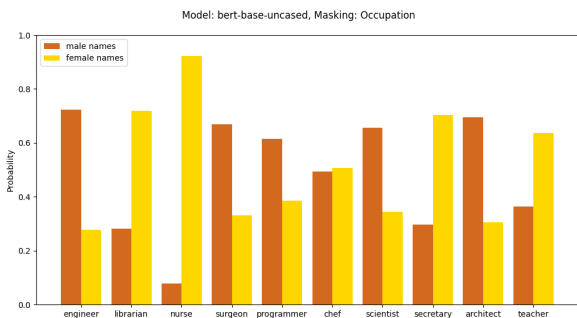


Figure 6.40: Original BERT Base Mask Occ exp 3

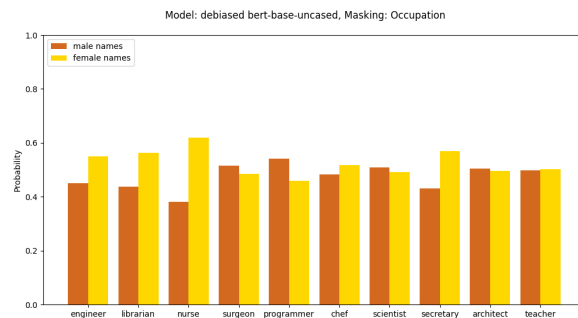


Figure 6.41: Debiased BERT Base Mask Occ exp 3

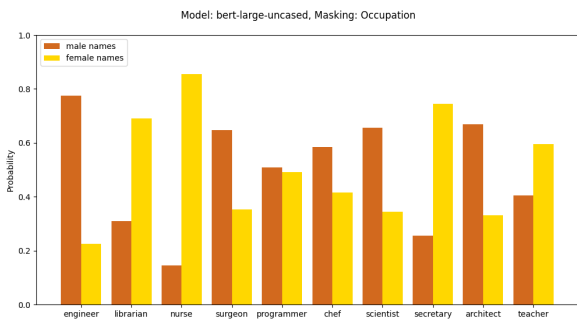


Figure 6.42: Original BERT Large Mask Occ exp 3

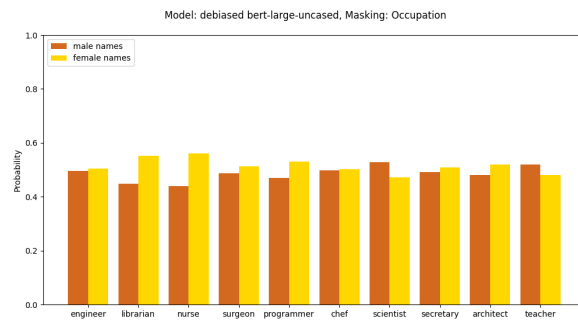


Figure 6.43: Debiased BERT Large Mask Occ exp 3

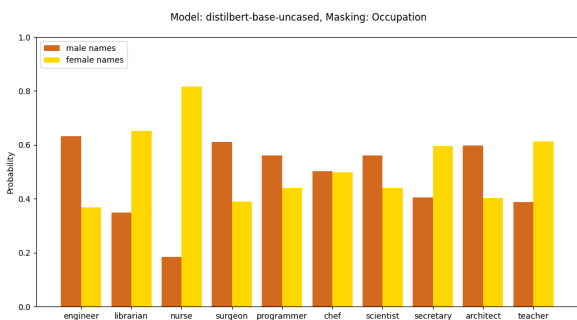


Figure 6.44: Original DistilBERT Mask Occ exp 3

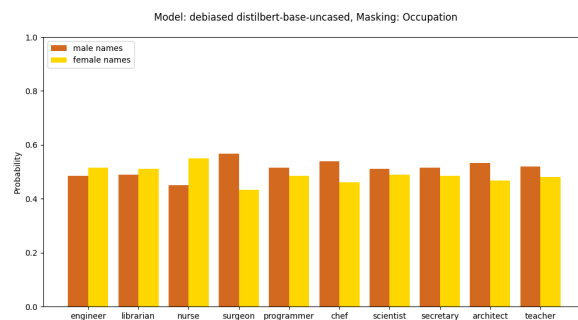


Figure 6.45: Debiased DistilBERT Mask Occ exp 3

After debiasing, the probability bar charts are more equalized. The gap between the probabilities of male and female is much reduced which was our primary aim. The models no longer favour one gender over the other when it comes to associating a particular occupation with that gender. However, we observed that after debiasing, in few of the graphs, the probabilities got reversed. For example, the probability of a male term or female term for a certain occupation was greater before debiasing but becomes less after debiasing. This could be happening due to the model overcorrecting the bias due to training on a small dataset. This happened in few cases and we will be looking into this issue in the future.

6.4 Quality Assurance of debiased BERT

We want our debiased models to be able to be used for downstream tasks. To ensure that the quality of BERT is retained after debiasing, we performed the Sentiment Analysis task from GLUE’s SST-2 (Stanford Sentiment Treebank) task [33]. GLUE is a framework for benchmarking and evaluating natural language understanding (NLU) models. The SST-2 is favoured because it provides a simple means of assessing the success of sentiment analysis without introducing complex linguistic or reasoning requirements. SST-2 contains IMDB movie review sentences utilised for sentiment analysis. The goal of SST-2 is to classify statements as positive or negative. The dataset contains 70k sentences where train set contains 67.3k sentences, test set contains 1.82k sentences and validation set contains 872 sentences. We kept batch size as 32, learning rate as 2×10^{-5} and number of epochs as 3 which was set as default in “run_glue.py”. We performed this task on both original biased BERT and debiased BERT and compared the accuracies. We found our debiased model to perform almost equally to the original model. To prove this, we performed paired t-test. The paired t-test is a method used to test whether the mean difference between pairs of measurements is zero or not. We debiased BERT Base according to exp 1: “he-she” 10 times with different seeds and performed SST-2 on each.

SST-2	Accuracy using original BERT	Accuracy using debiased BERT
1	0.9232	0.9186
2	0.9232	0.9197
3	0.9232	0.9243
4	0.9232	0.9300
5	0.9232	0.9255
6	0.9232	0.9232
7	0.9232	0.9278
8	0.9232	0.9255
9	0.9232	0.9220
10	0.9232	0.922

Table 6.5: Before and After Comparison of SST-2 Accuracy

Null Hypothesis (H_0): The mean accuracy of debiased BERT is not significantly different from the mean accuracy of the original BERT on the SST-2 task.

Alternative Hypothesis (H_a): The mean accuracy of debiased BERT is significantly different from the mean accuracy of the original BERT on the SST-2 task.

$$\text{Mean } (\mu) = \frac{\sum \text{Differences}}{n} = -0.00046$$

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{\sum (\text{Differences} - \text{Mean})^2}{n-1}} \approx 0.00336$$

$$\text{t-statistic } (t) = \frac{\text{Mean}}{\frac{\sigma}{\sqrt{n}}} \approx -0.463$$

$$p\text{-value} = 0.5671627142$$

In this case, since the p -value (0.5671627142) is greater than the significance level ($\alpha = 0.05$), we fail to reject the null hypothesis. There is no statistically significant difference between the “before” and “after” values at the 0.05 significance level.

Chapter 7

Conclusion

Similar to static word embedding models, contextualized word embeddings are also prone to sexism. We examined gender bias in BERT and other transformer models and discovered that the training corpus contains a sizeable gender bias dataset. Our work provides the foundation for evaluating and mitigating bias in downstream applications, which is especially important as contextualized embeddings are increasingly used to improve performance on key NLP tasks like BERT [23]. The goal of this work is to analyze the gender bias in models that produce contextualized word embeddings like BERT, ALBERT, RoBERTa, and DistilBERT and mitigate the bias so that these models do not carry forward the bias to the downstream tasks. For measuring gender bias, we tried using cosine similarity, using direct bias test following Basta [20], and finally, we showed bias using masked probability, taking ideas from Kurita et al. [25], but in our own way. For debiasing the models, we applied CDA to datasets to create a gender balanced dataset and continued pretraining the models on these datasets.

7.1 Findings and Contributions

Bias in these models are dangerous and often overlooked. There is not enough work on it and it is still an ongoing research. We found out that gender bias in contextualized word embeddings cannot be detected through traditional methods like cosine similarity which works on static word embeddings. We also found that the different transformer models have different vocabulary or tokenization process for which few models such as RoBERTa and ALBERT do not recognize all words such as the occupations and male/female words we used. Our main contributions are that we firstly analyzed the existing bias in different models thoroughly through masking probability. We employed two techniques - masking the gendered term and masking the occupation to analyze the bias in both direction. We then created an evaluation metric - MALoR that can be applied to any of the models that gives us an idea of how much bias a model contains based on gendered pronouns and gendered male and female names. For the evaluation metric, we created a wide range of sentence structures for the three different experiments we did - “he-she”, “his-her” and “male-female name” so that the bias representation is more clear. Next, for debiasing the models, we applied CDA to datasets like news corpus and news commentary to make it gender balanced and continued pretraining the models until the learning curve converges. Finally, we compared the gender bias of the

different models before and after training and showed our method is working by applying it on 3 different experiments.

7.2 Limitations and Future Work

Since, we only worked on gendered pronouns - “he-she”, “his-her” and gendered names - “male name-female name”, we plan to work on other gender nouns like “father-mother”, “boy-girl”, etc so that the model is completely gender bias free. On top of that, since our debiasing method is a time consuming process which needs to be applied to each of the models individually, we lacked the time and resource to train all the models. Hence, we plan to continue the process to each of the models to create a detailed comparison. Since RoBERTa and ALBERT do not contain the occupational words that we have used throughout the experiments and the male and female names in their vocabulary, we will work in the future to overcome these issues. Lastly, we will look into the issue where the probability bar charts were getting reversed before and after debiasing in few cases and we will also work to achieve a smoother convergence in our learning graphs.

Bibliography

- [1] W. Martyna, “What does ‘he’ mean?” *Journal of Communication*, vol. 28, no. 1, pp. 131–138, 1978. DOI: <https://doi.org/10.1111/j.1460-2466.1978.tb01576.x>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-2466.1978.tb01576.x>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.1978.tb01576.x>.
- [2] A. G. Greenwald, D. E. McGhee, and J. L. K. Schwartz, “Measuring individual differences in implicit cognition: The implicit association test,” *Journal of personality and social psychology*, vol. 74, pp. 1464–1480, 1998.
- [3] *EMNLP-CoNLL ’12: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea: Association for Computational Linguistics, 2012.
- [4] B. Schmidt, “Rejecting the gender binary: A vector-space operation,” Oct. 2015. [Online]. Available: <http://bookworm.benschmidt.org/posts/2015-10-30-rejecting-the-gender-binary.html>.
- [5] W. on Statistical Machine Translation, *News 2014 shuffled v2 dataset for wmt15*, <https://data.statmt.org/wmt15/news.2014.en.shuffled.v2.gz>, Accessed: yyyy-mm-dd, 2015.
- [6] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” *NIPS*, 2016. DOI: 10.48550/arXiv.1607.06520. eprint: <https://doi.org/10.48550/arXiv.1607.06520>. [Online]. Available: <https://doi.org/10.48550/arXiv.1607.06520>.
- [7] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” *Science*, vol. 356, no. 6334, pp. 183–186, 2017.
- [8] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Men also like shopping: Reducing gender bias amplification using corpus-level constraints,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2979–2989. DOI: 10.18653/v1/D17-1323. [Online]. Available: <https://aclanthology.org/D17-1323>.
- [9] 2018. [Online]. Available: <https://www.ssa.gov/oact/babynames/decades/century.html>.
- [10] J. Alammam, *The illustrated transformer*, 2018. [Online]. Available: <https://jalammam.github.io/illustrated-transformer/>.

- [11] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. S. Zemel, “Understanding the origins of bias in word embeddings,” *ArXiv*, vol. abs/1810.03611, 2018.
- [12] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, vol. 81, 2018, pp. 77–91. [Online]. Available: <https://proceedings.mlr.press/v81/buolamwini18a.html>.
- [13] S. Kiritchenko and S. Mohammad, “Examining gender and race bias in two hundred sentiment analysis systems,” in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 43–53. DOI: 10.18653/v1/S18-2005. [Online]. Available: <https://aclanthology.org/S18-2005>.
- [14] M. E. Peters, M. Neumann, M. Iyyer, *et al.*, “Deep contextualized word representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. [Online]. Available: <https://aclanthology.org/N18-1202>.
- [15] M. E. Peters, M. Neumann, L. Zettlemoyer, and W. Yih, “Dissecting contextual word embeddings: Architecture and representation,” *CoRR*, vol. abs/1808.08949, 2018. arXiv: 1808.08949. [Online]. Available: <http://arxiv.org/abs/1808.08949>.
- [16] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, “Gender bias in coreference resolution,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018.
- [17] W. on Statistical Machine Translation, *News commentary v13 dataset for wmt18*, <https://data.statmt.org/wmt18/translation-task/news-commentary-v13.en.gz>, Accessed: yyyy-mm-dd, 2018.
- [18] K. Webster, M. Recasens, V. Axelrod, and J. Baldrige, “Mind the GAP: A balanced corpus of gendered ambiguous pronouns,” *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 605–617, 2018. DOI: 10.1162/tacl_a_00240. [Online]. Available: <https://aclanthology.org/Q18-1042>.
- [19] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, “Learning gender-neutral word embeddings,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4847–4853. DOI: 10.18653/v1/D18-1521. [Online]. Available: <https://aclanthology.org/D18-1521>.
- [20] C. Basta, M. R. Costa-jussà, and N. Casas, “Evaluating the underlying gender bias in contextualized word embeddings,” in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 33–39. DOI: 10.18653/v1/W19-3805. [Online]. Available: <https://aclanthology.org/W19-3805>.

- [21] K. Chaloner and A. Maldonado, “Measuring gender bias in word embeddings across domains and discovering new gender bias word categories,” in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 25–32. DOI: 10.18653/v1/W19-3804. [Online]. Available: <https://aclanthology.org/W19-3804>.
- [22] S. Dev and J. Phillips, “Attenuating bias in word vectors,” in *International Conference on Artificial Intelligence and Statistics*, 2019. DOI: 10.48550/ARXIV.1901.07656. [Online]. Available: <https://arxiv.org/abs/1901.07656>.
- [23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: <https://aclanthology.org/N19-1423>.
- [24] H. Gonen and Y. Goldberg, “Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them,” *arXiv preprint arXiv:1903.03862*, 2019.
- [25] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, *Measuring bias in contextualized word representations*, 2019. arXiv: 1906.07337 [cs.CL].
- [26] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A lite BERT for self-supervised learning of language representations,” *CoRR*, vol. abs/1909.11942, 2019. arXiv: 1909.11942. [Online]. Available: <https://dblp.org/rec/journals/corr/abs-1909-11942.bib>.
- [27] Y. Liu, M. Ott, N. Goyal, *et al.*, “Roberta: A robustly optimized BERT pre-training approach,” *CoRR*, vol. abs/1907.11692, 2019. arXiv: 1907.11692. [Online]. Available: <http://arxiv.org/abs/1907.11692>.
- [28] R. H. Maudslay, H. Gonen, R. Cotterell, and S. Teufel, “It’s all in the name: Mitigating gender bias with name-based counterfactual data substitution,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5267–5275. DOI: 10.18653/v1/D19-1530. [Online]. Available: <https://aclanthology.org/D19-1530>.
- [29] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, “On measuring social biases in sentence encoders,” *CoRR*, vol. abs/1903.10561, 2019. arXiv: 1903.10561. [Online]. Available: <http://arxiv.org/abs/1903.10561>.
- [30] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter,” *ArXiv*, vol. abs/1910.01108, 2019.
- [31] D. Shah, H. A. Schwartz, and D. Hovy, “Predictive biases in natural language processing models: A conceptual framework and overview,” *CoRR*, vol. abs/1912.11078, 2019. arXiv: 1912.11078. [Online]. Available: <http://arxiv.org/abs/1912.11078>.
- [32] T. Sun, A. Gaut, S. Tang, *et al.*, “Mitigating gender bias in natural language processing: Literature review,” *arXiv preprint arXiv:1906.08976*, 2019.

- [33] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, *Glue: A multi-task benchmark and analysis platform for natural language understanding*, 2019. arXiv: 1804.07461 [cs.CL].
- [34] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, “Gender bias in contextualized word embeddings,” *arXiv preprint arXiv:1904.03310*, 2019.
- [35] M. Bartl, M. Nissim, and A. Gatt, “Unmasking contextual stereotypes: Measuring and mitigating bert’s gender bias,” *CoRR*, vol. abs/2010.14534, 2020. arXiv: 2010.14534. [Online]. Available: <https://arxiv.org/abs/2010.14534>.
- [36] S. Dev, T. Li, J. M. Phillips, and V. Srikumar, “On measuring and mitigating biased inferences of word embeddings,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 7659–7666.
- [37] S. Gururangan, A. Marasovic, S. Swayamdipta, *et al.*, “Don’t stop pretraining: Adapt language models to domains and tasks,” *CoRR*, vol. abs/2004.10964, 2020. arXiv: 2004.10964. [Online]. Available: <https://arxiv.org/abs/2004.10964>.
- [38] V. Kumar, T. S. Bhotia, V. Kumar, and T. Chakraborty, “Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 486–503, 2020. DOI: 10.1162/tacl_a_00327. [Online]. Available: <https://aclanthology.org/2020.tacl-1.32>.
- [39] T. Wang, X. V. Lin, N. F. Rajani, B. McCann, V. Ordonez, and C. Xiong, “Double-hard debias: Tailoring word embeddings for gender bias mitigation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 2020, pp. 5443–5453. DOI: 10.18653/v1/2020.acl-main.484. [Online]. Available: <https://aclanthology.org/2020.acl-main.484>.
- [40] T. Wolf, L. Debut, V. Sanh, *et al.*, “Transformers: State-of-the-Art Natural Language Processing,” Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- [41] B. Kumar. “Bert variants and their differences - 360digitmg.” Accessed: May 20, 2023. (2021), [Online]. Available: <https://360digitmg.com/blog/bert-variants-and-their-differences>.
- [42] Q. Liu, M. Kusner, and P. Blunsom, “Counterfactual data augmentation for neural machine translation,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 187–197. DOI: 10.18653/v1/2021.naacl-main.18. [Online]. Available: <https://aclanthology.org/2021.naacl-main.18>.
- [43] N. T. Lee, P. Resnick, and G. Barton, “Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms,” *Brookings*, 2022, Published on May 22, 2022. [Online]. Available: <https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>.

- [44] B. Muller. “Bert 101 - state of the art nlp model explained.” (Mar. 2022), [Online]. Available: <https://huggingface.co/blog/bert-101>.
- [45] P. Nemani, Y. D. Joel, P. Vijay, and F. F. Liza, *Gender bias in transformer models: A comprehensive survey*, 2023. arXiv: 2306.10530 [cs.CL].