

Kennesaw State University

DigitalCommons@Kennesaw State University

---

Analytics and Data Science Dissertations

Ph.D. in Analytics and Data Science Research  
Collections

---

Spring 4-28-2023

## Quantification of Various Types of Biases in Large Language Models

Sudhashree Sayenju

Follow this and additional works at: [https://digitalcommons.kennesaw.edu/dataphd\\_etd](https://digitalcommons.kennesaw.edu/dataphd_etd)



Part of the [Artificial Intelligence and Robotics Commons](#), [Data Science Commons](#), [Numerical Analysis and Scientific Computing Commons](#), [Risk Analysis Commons](#), and the [Statistical Models Commons](#)

---

### Recommended Citation

Sayenju, Sudhashree, "Quantification of Various Types of Biases in Large Language Models" (2023). *Analytics and Data Science Dissertations*. 16.  
[https://digitalcommons.kennesaw.edu/dataphd\\_etd/16](https://digitalcommons.kennesaw.edu/dataphd_etd/16)

This Dissertation is brought to you for free and open access by the Ph.D. in Analytics and Data Science Research Collections at DigitalCommons@Kennesaw State University. It has been accepted for inclusion in Analytics and Data Science Dissertations by an authorized administrator of DigitalCommons@Kennesaw State University. For more information, please contact [digitalcommons@kennesaw.edu](mailto:digitalcommons@kennesaw.edu).

# Quantification of Various Types of Biases in Large Language Models

**Sudhashree Sayenju**

**Chair:**

Dr. Ramazan Aygun

**Committee members:**

Dr. Ying Xie

Dr. Yifan Zhang

A Dissertation Presented for the Doctor of Philosophy

Degree at



April 2023

---

© by Sudhashree Sayenju, 2023

All Rights Reserved.

# Acknowledgements

My journey to pursue a PhD began as a dream when I was working as a software developer in Germany. I feel blessed and lucky that I got to follow my dreams and land here in the US. I want to thank every professor, colleague and classmate I got a chance to collaborate with. I also want to thank some of my family members and friends who showed me support in other ways.

My supervisor, Dr. Ramazan Aygun has always given me the guidance, encouragement and structure I needed during my research. Whether it be in the simplest things asking me to make a list of potential conferences to submit my paper to, or technical advice and instructions I received, I am grateful for his time, contribution and faith in me. I would also like to thank my committee member Dr. Ying Xie and Dr. Yifan Zhang for providing me perspectives and input that I believe have contributed in improving my research. Our School of Data Science and Analytics' previous executive director, Dr. Jennifer Priestley has always made me feel like all she wants to see is our success. Other than academically, she also helped me when I had living situation trouble in the beginning of COVID-19 pandemic. Thank you Dr. Priestley for being an angel. Our current program director Dr. Sherrill Hayes is no different. Thank you Dr. Hayes for checking in with me and making sure my troubles and concerns are always heard of. I also want to thank him for conversations about bias in data based systems. Our Associate Program Director, Dr. Sherry Ni made sure to monitor my progress throughout this journey. She has also always encouraged and celebrated every milestone I achieved during my doctoral studies. I would also like to thank Professor Bill Franks for giving feedback, presentation tips, encouragement and fun conversations.

---

As a part of our Data Science Lab program, I got the opportunity to collaborate with Dr. Sereres Johnston, George Lee, Dr. Girish Modgil, Dan Sullivan and Dr. Hansook Choi from Travelers. I would like to thank them for their cooperation, feedback, support and encouragement.

I was fortunate to make wonderful friends over the course of my PhD studies. Thank you Christina Stradwick for being a fantastic friend and roommate. When I first came to the US, I had a difficult living situation. I was relieved when she asked me to be her roommate. I finally had piece of mind when I was home. The kindness that Dr. Lili Zhang and Dr. Seema Sangari in my earliest days of the program cannot be forgotten either. Thank you Nina Grundlingh, Andrew Henshaw, Jitendra Sai Kota, Yihong Zhang, Kate Mobley, Sahar Yarmohammadtoosky, Dr. Lauren Staples, Sanjoosh Akkineni, Vanessa Sunar, Dr. Yan Wang, Dr. Trent Geisler, Srivatsa Mallapragada, Dr. Jonathan Boardman, Duleep Prasanna Rathgamage Don and Mallika Boyapati for wonderful memories, fun times or project collaborations. Friends I made in Germany, namely Dr. Maria Kalweit, Irina Chernigova, Zaher Wanli, Benjamin Senyoni and Jan Bechtold have also been supportive and encouraging throughout my PhD journey albeit virtually.

This acknowledgement section would be incomplete without thanking the reason for my existence, my parents and grandparents. Thank you for always pushing me to get the best of the opportunities available to me. I would also like to thank my cousins, aunts and uncles who have been there with me through happy and difficult times.

This work was supported primarily by the Travelers Indemnity Company. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the Travelers Indemnity Company.

# Quantification of Various Types of Biases in Large Language Models

## Abstract

Natural Language Processing (NLP) systems are included everywhere on the internet from search engines, language translations to more advanced systems like voice assistant and customer service. Since humans are always on the receiving end of NLP technologies, it is very important to analyze whether or not the Large Language Models (LLMs) in use have bias and are therefore unfair. The majority of the research in NLP bias has focused on societal stereotype biases embedded in LLMs. However, our research focuses on all types of biases, namely model class level bias, stereotype bias and domain bias present in LLMs. Model class level bias happens when a model tends to favor some classification labels or outputs compared to the others. We investigate how a classification model hugely favors one class with respect to another. We propose a bias evaluation technique called *directional pairwise class confusion bias* that highlights an LLM's bias on pairs of classes. Unfavorable kind of stereotype bias takes place when LLMs cause significant injustice or harm to disadvantaged or marginalized group of people. Although the most advanced deep LLMs claim to mimic human responses via powerful and sophisticated algorithms, the capabilities that such models offer have shown to possess bias. Quantifying such stereotype biases appropriately is essential so that the bias measures can be used to calibrate potential harm the models can cause. On the other hand, domain biases are desired for the model because it indicates the model is learning necessary facts for it to be powerful. We devise techniques to measure class level, stereotype, and domain biases appropriately.

# Contents

<b>1</b>	<b>Introduction</b>	<b>13</b>
1.1	Bias in NLP Models . . . . .	14
1.2	Importance of Bias Measures . . . . .	16
1.3	Motivation . . . . .	17
1.4	Approach . . . . .	18
1.4.1	Detect Model Class Level Bias . . . . .	19
1.4.2	Quantify Unfavorable Stereotype Bias . . . . .	20
1.4.3	Quantify Domain Knowledge . . . . .	21
1.5	Copyrights and Permissions . . . . .	21
1.6	Dissertation Organization . . . . .	23
<b>2</b>	<b>Related Work</b>	<b>24</b>
2.1	Types of Biases . . . . .	24
2.2	Class Level Bias . . . . .	27
2.3	Unfavorable Stereotype Bias in LLMs . . . . .	28
2.3.1	Word2vec . . . . .	29
2.3.2	GloVe . . . . .	30
2.3.3	CBOW . . . . .	34
2.3.4	ELMo . . . . .	35
2.3.5	BERT . . . . .	38
2.3.6	GPT Models . . . . .	40
	GPT . . . . .	40

GPT-2 . . . . .	42
GPT-3 . . . . .	43
GPT-4 . . . . .	44
ChatGPT . . . . .	45
2.3.7 Summary of Stereotype Biases in LLMs . . . . .	46
2.4 Bias Measures . . . . .	48
2.4.1 WEAT and its Derivatives . . . . .	48
2.4.2 SAME . . . . .	49
2.4.3 Limitations of WEAT and SAME . . . . .	50
2.5 Quantify Model Tendency after Customization . . . . .	51
2.6 Summary . . . . .	52
<b>3 Directional Pairwise Class Confusion Bias for Evaluating Class Level Bias</b>	<b>54</b>
3.1 Motivation . . . . .	55
3.2 Building Intent Classification Model using Transfer Learning with BERT Model . . . . .	56
3.3 Directional Pairwise Class Confusion Bias . . . . .	57
3.3.1 Definition: Directional Pairwise Class Confusion Bias . . . . .	61
3.3.2 Bias Mitigation Process . . . . .	63
3.4 Experiments . . . . .	64
3.4.1 Data . . . . .	64
3.4.2 Bias in the Original BERT Model for Chatbot’s Intent Classification	65
3.4.3 Priori Bias Mitigator . . . . .	66
3.4.4 Posteriori Bias Mitigator . . . . .	66
3.4.5 Discussion . . . . .	67
3.5 Summary . . . . .	68
<b>4 Differential Cosine Bias Measure for Evaluating Stereotype and Categorical Bias</b>	<b>70</b>
4.1 Motivation . . . . .	71



4.1.1	Definition of Sets . . . . .	73
4.2	Differential Cosine Bias (DiCoBi) Measure . . . . .	73
4.2.1	DiCoBi on Singleton Sets . . . . .	73
4.2.2	General DiCoBi Measure . . . . .	74
4.3	Experiments . . . . .	77
4.3.1	Word Groups . . . . .	77
4.3.2	NLP Model for Bias Analysis . . . . .	78
4.3.3	Bias Validation . . . . .	79
4.3.4	Gender Bias . . . . .	79
4.3.5	Word Group Level Experiments . . . . .	80
4.4	Discussion . . . . .	81
4.5	Summary . . . . .	82
<b>5</b>	<b>Quantifying Domain Knowledge for Evaluating Domain Bias</b>	<b>84</b>
5.1	Motivation . . . . .	85
5.2	Domain Knowledge . . . . .	85
5.3	Model Tendency Visualization . . . . .	86
5.3.1	Proposed Method for Selecting $\alpha_-$ and $\alpha_+$ . . . . .	87
5.3.2	Arrows and their implications . . . . .	87
5.4	Experiments . . . . .	89
5.5	Summary . . . . .	90
<b>6</b>	<b>Conclusions and Future Work</b>	<b>92</b>
6.1	Future Work . . . . .	93

# List of Figures

1.1	Within an insurance domain, we want <i>Umbrella</i> to mean the type of insurance coverage instead of device used to protect from rain or heat . . .	22
2.1	Summary of bias analyses done in Word2Vec based on published papers. . .	31
2.2	Summary of bias analyses done in GloVe based on published papers. . . .	33
2.3	Summary of bias analyses done in CBOW through based on published papers. . . . .	36
2.4	Summary of bias analyses done in ELMo based on published papers. . . .	37
2.5	Summary of bias analyses done in BERT based on published papers. . . .	41
2.6	Summary of bias analyses done in GPT-2 through based on published papers. . . . .	44
2.7	Example of an inquiry to ChatGPT ( <a href="https://chat.openai.com/">https://chat.openai.com/</a> ) and its response. . . . .	46
2.8	Graph of complexity of models over time for bias analyses based on published papers. . . . .	47
2.9	Graph of various bias analyses for the popular LLMs based on published papers. . . . .	47
3.1	Results of all 4 epochs in the training phase. © 2022 IEEE . . . . .	57
3.2	Class Confusion matrix © 2022 IEEE . . . . .	59
3.3	Heatmap of Class Confusion matrix © 2022 IEEE . . . . .	60
3.4	Dividing each cell in the confusion matrix by the maximum of its column. © 2022 IEEE . . . . .	60

3.5 Directional Pairwise Class Confusion Bias © 2022 IEEE . . . . . 62

3.6 Pruned Directional Pairwise Class Confusion Bias matrix after threshold  
was set to 0.15. © 2022 IEEE . . . . . 62

3.7 Evaluation of bias mitigation process © 2022 IEEE . . . . . 64

3.8 Test set evaluation on original intent classification BERT model without  
mitigating bias. . . . . 65

4.1 Difference vectors of gender and occupation are parallel  $\cos(\overrightarrow{man-woman}, \overrightarrow{doctor-nurse}) = 1$  . . . . . 75

4.2 Difference vectors of gender and occupation are parallel . . . . . 75

4.3 Cases where our differential cosine bias metric  $\cos(\overrightarrow{A1} - \overrightarrow{B1}, \overrightarrow{A2} - \overrightarrow{B2}) = 0$  75

5.1 Range of *domain\_gain* . . . . . 87

# List of Tables

2.1	Papers analyzing bias in Word2vec. . . . .	30
2.2	Papers analyzing bias in GloVe. . . . .	32
2.3	Papers analyzing bias in CBOW. . . . .	34
2.4	Papers analyzing bias in ELMo. . . . .	37
2.5	Papers analyzing bias in BERT. . . . .	39
2.6	Papers analyzing bias in GPT. . . . .	41
2.7	Papers analyzing bias in GPT-2. . . . .	42
2.8	Papers analyzing bias in GPT-3. . . . .	45
2.9	Bias measures similar to WEAT . . . . .	49
2.10	Quantifying categorical biases using WEAT and SAME . . . . .	51
2.11	Quantifying gender bias using WEAT and SAME . . . . .	51
3.1	Priori Bias Mitigation: Before (original BERT model) and after ( $c_d$ for training secondary model) performance of source classes $c_{coverage}$ , $c_{billing}$ , $c_{everythingElse}$ . Improvement is seen in the recall of source classes $c_{coverage}$ and $c_{billing}$ (in bold font). . . . .	66
3.2	Posteriori Bias Mitigation: Before (original BERT model) and after ( $c_d$ for training secondary model) performance of source classes in biased pairs. Largest Recall for each class are written in bold font. . . . .	67
3.3	Priori and Posteriori Bias Mitigation: Performance of source classes in biased pairs on all experiments. Largest Recall for each class are written in bold font. . . . .	68

4.1	Word sets to test various types of biases . . . . .	78
4.2	Quantifying categorical biases for word level experiments . . . . .	80
4.3	Quantifying gender bias for word level experiments . . . . .	80
4.4	Quantifying various types of biases for word group level experiments . . . . .	81
5.1	Indications for values of average difference in magnitude . . . . .	86
5.2	Description and tendencies of same zone arrows. . . . .	88
5.3	Examples of some transition zone arrows with their description and tendencies. . . . .	88
5.4	The <i>domain_gain</i> for various tests in models <i>Model_General</i> and <i>Model_Domain</i> . . . . .	90
5.5	Knowledge gain summary of models . . . . .	91

# Chapter 1

## Introduction

Natural Language Processing (NLP) is the task of processing human language with the use of computers. In the beginning of 1950s, NLP involved complex sets of hand written rules [1, 2]. The use of statistical models in NLP tasks only began in the early 1990s ([3, 4, 5]). Statistical models were first used for language translation tasks [6, 7]. Today NLP is used to solve multiple tasks such as Tokenization, Part-of-speech tagging, Dependency Parsing, Lemmatization, Stemming, Stopword Removal, Named Entity Recognition (NER), Classification, and more. As computing resources became available, machine learning algorithms were deployed. Over time, the complexity of machine learning algorithms have increased and become more powerful for various NLP tasks. Simultaneously, there was also the rise of World Wide Web and digital text data.

Simple machine learning algorithms used TF-IDF (Term Frequency-Inverse Document Frequency) to detect language patterns prior to the rise of deep neural networks in NLP [8]. In 2017, transformer based models [9] gave rise to Large Language Models (LLMs) with billions or more trainable parameters. In the past decade, LLMs have made breakthroughs for various NLP tasks. Recently popular models like BERT [10], GPT-2 [11], GPT-3 [12], and GPT-4 [13] are trained on enormous text corpus and fine-tuned for various specific tasks. Although the performance of these models appear to be very close to that of a human, it is essential to be aware of various kinds of biases in these advanced NLP models. We should be cognisant of the fact that historical texts contain various kinds of biases. Therefore, LLMs are excellent not only in learning language

patterns but also in learning the biases present in those texts.

NLP is a field that always involves humans at consuming end of the models or in the decision making process. Therefore, to ensure fairness, a model should be analyzed for *stereotype bias*. It should be noted that stereotype bias can be favorable as well as unfavorable. For example, women being attached to child care terms makes sense due to biological reasons and is therefore favorable gender bias, whereas women being linked to low IQ jobs is unfavorable gender bias. Unfavorable stereotype bias can appear at various stages of modelling including data collection, training, or deploying the model for a completely different purpose. On the other hand, we should be aware of the fact that some biases which humans consider as general knowledge is desired in the NLP models. We refer these necessary biases as *categorical bias* throughout this document. Another type of favorable bias is *domain bias*. When language models show semantic tendencies towards their domain interpretations for words with complete different meanings in layman sense, we define it to have domain bias. When domain biased models are deployed within their applications, the interpretations of polysemic words are interpreted as intended within their domain. While the presence of categorical bias shows general knowledge and facts were learned, domain bias shows the correct meaning of polysemic words were recognized for their domain.

## 1.1 Bias in NLP Models

Naturally, the text corpora produced by humans of the society will not be free of prejudices. Machine learning models that learn from this data will generally reflect these biases unless mitigation strategies are applied. The data used in NLP models are either collected via surveys/experiments or simply text corpora like Wikipedia [14], Reddit [15], Google News [16] and more. Bias in data collection could originate from low representation of demographically underrepresented groups. On the other hand, text corpus such as Wikipedia contains huge amount of historical information written by or is about white male scientists, philosophers, litterateurs, politicians, etc. As the culture evolves over years, what was an acceptable norm 50 years ago may not be viable today. These

Wikipedia pages could potentially embed systemic racism and other forms of discrimination due to historical context and writers. Additionally, the Reddit Corpus includes a lot of customer comment threads which may include abusive or toxic language. Offensive language in data can also hand over potential bias in a model. The Google News text corpus has shown incidents of racism, gender discrimination and others when used to train language models [17, 18]. Therefore, the data itself represents the unfavorable stereotype biases and discrimination in the society. It is not surprising that biases in data are inherited onto later stages of modelling.

Ultimately, NLP models are deployed in real-world systems such as chatbots, translation, search engines, document-classification, text annotation, sentiment analysis, etc. and eventually impact decision making. Decisions made out of these systems affect crucial aspects of life such as education, health, work competence, job opportunities, commute, interest rates, and self-conduct. While using LLMs, one needs to be mindful of these decisions as it affects the quality of life for humans belonging to some stereotyped minority. Since decisions produced by unfavorably biased models could be regarded as unacceptable or unfair depending on the context, it is critical to understand the relationship between the bias in data and bias in algorithms. The machine learning models are capable of learning the bias in data during training and also reflect the bias in their predictions. Additionally, algorithms can alter the level of bias in data or display a bias that does not exist in data. In cases where reinforcement learning is added to the system to improve the model, it may produce even more biased data for future model training [19].

The more sophisticated a LLM is, the more vulnerable it is to biases as it can learn peculiarities in the data with the goal of increasing accuracy or minimizing loss function ignoring biases. Bias in a model can arise from two situations. Firstly, as machine learning models are data driven, the bias in the data might be so prominent that the model acquires the same bias or worse amplifies it. Secondly, the complex and non-interpretable mechanism in deep neural network algorithms can exhibit bias that is not present in the data [20]. If such unfavorably biased models are deployed in real-world systems, it could consequently affect marginalized communities [19].



The word *bias* in data or a predictive model mostly refers to unfavorable stereotype bias in majority of research in this field. Unfavorable stereotype bias can be identified through structured or semi-structured data in the form of protected attributes [21, 22, 23, 24]. Popular tools like AI Fairness 360 (AIF360) [25] and AWS Sagemaker Clarify [26] address bias related to protected attributes such as age, gender, race, ethnicity and more. These biases found in protected attributes are part of the input features. In unstructured data like text, stereotyping bias can also culminate into semantic biases. Additionally, deviations from standard text corpora like grammatical errors, spelling mistakes, accents and regional lingo might be embedded in the semantics of the text. The accumulation of all these biases is capable of influencing model behavior and classification label preferences. While there is research that solely addresses the class imbalance problem [27, 28, 29, 30], not much research has been conducted in model’s classification bias due to various anomalies enclosed in the text semantics. A model could favor a certain class more than another in a task like intent classification. Therefore, defining and quantifying such *class-specific bias* can help find the cause of the bias and eventually build procedures to mitigate it.

## 1.2 Importance of Bias Measures

Currently, bias evaluation measures are used as a diagnostic technique on fully trained models [31, 32, 33, 34]. Instead, bias measures can be used as a regularization method by either monitoring during the training epochs or incorporated into the loss function of the model. In doing so, we could make models less harmful or more beneficial in their long term applications.

When models are built, multiple variations with altered hyper-parameters or datasets are usually tested. When choosing the best model, the bias measure values need to be taken into account along with the standard performance metrics like accuracy, precision, recall, F1-score and more. Choosing the best model should not just depend on the bias measure value but also the type of bias being monitored.

While some biases are deemed favorable, depending on the context the same type of

bias could be unfavorable. For unfavorable biases, such as class level bias, investigating the bias measure values is the first step towards devising mitigation techniques. Stereotype bias is generally considered unfavorable in the literature with some exceptions. Bias measures should be investigated before and after mitigation to make sure the mitigation works. On the other hand, favorable kinds of bias such as obtaining domain knowledge need to be observed using bias evaluation measures to check whether or not transfer learning improves customization in models.

### 1.3 Motivation

While biases can be integrated in the model at various stages, those biases might also be of various types. When we think of the word *bias*, we tend to associate it with its negative connotations. However, there are good kinds of biases that are necessary as well. Quantifying the favorable and unfavorable biases is necessary to take future steps such as mitigating the bad types of bias and enhance the good types of bias. Simply put, quantification of bias is a step towards making LLMs fairer. Therefore, in this dissertation we quantify various kinds of biases namely: class level bias, unfavorable stereotype bias and domain bias.

Due to various reasons like class imbalance, semantic noise in texts and insufficient training data, the performance of a model may not be high. The accuracy of a model provides a rough performance indication while hiding the actual model performance for imbalanced datasets. On the other hand, measures such as sensitivity, specificity, precision, recall, and F1-measure yield class level performance. None of these measures points out the model's bias for favoring one class over another class, which will be referred as *class level bias* in this dissertation. As such class level bias is not studied in the literature, we will be defining and quantifying such bias for classification models.

In the field of quantifying unfavorable stereotype bias, we found out via multiple experiments that well-known bias evaluation metrics like WEAT [35] and SAME [36] do not always measure bias accurately. Some results indicated that the metrics measure the co-occurrence of words or terms that they learned from the text corpus they were trained

on rather than the stereotype bias. Another, drawback of these measures is that they are not comparable. In other words, if the values for a bias measure of one example of gender bias is higher than that of a second example, it does not necessarily mean the first example is more biased than the second. To address this issue, we introduce a novel bias evaluation called *Differential Cosine Bias Measure*. We show that our measure works by comparing the unwanted bias that the model captures against on categorical biases that the model should learn.

Transformer based Large language models such as BERT [10] have demonstrated the ability to derive contextual information from the words surrounding it. However, when these models are applied in specific domains such as medicine, insurance, or scientific disciplines, publicly available models trained on general knowledge sources such as Wikipedia, it may not be as effective in inferring the appropriate context compared to domain-specific models trained on specialized corpora. Given the limited availability of training data for specific domains, pre-trained models can be fine-tuned via transfer learning using relatively small domain-specific corpora. However, there is currently no standardized method for quantifying the effectiveness of these domain-specific models in acquiring the necessary domain knowledge. To address this issue, we explore hidden layer embeddings and introduce *domain\_gain*, a measure to quantify the ability of a model to infer the correct context. In this dissertation, we show how our measure could be utilized to determine whether words with multiple meanings are more likely to be associated with domain-related meanings rather than their colloquial meanings.

## 1.4 Approach

In terms of bias analysis of a LLM, the main question is whether the model is biased or not. Nevertheless, the presence of a bias is not a binary decision. Hence, measuring the degree of existing bias in a LLM is crucial because quantifying bias enables i) choosing a model that has the lowest unfavorable bias, ii) developing bias mitigation techniques and checking whether the bias is actually mitigated or not, iii) determining whether the bias is negligible or not, and iv) monitoring whether the model is able to learn the domain

knowledge necessary to be used for various domain specific tasks. As the main goal of this dissertation, we quantify three types of biases: class level bias, unfavorable stereotype bias, and domain bias. The contributions of this dissertation can be categorized into three as follows:

1. Define and quantify class level bias of an NLP model favoring one class to another class while predicting the class of an unseen data.
2. Quantify unfavorable stereotype bias that can impact minority or underrepresented groups while ensuring that the categorical bias (facts or knowledge) is maintained, and
3. Quantify domain bias for models that are trained for specific domains if the polysemic words have tendency to their domain specific meanings rather than their colloquial usage.

### 1.4.1 Detect Model Class Level Bias

The misclassification errors occurs due to a combination of class imbalance, semantic noise in texts and insufficient variety of topics in training text corpus. We delve into misclassification pairs, where each pair consists of the true label and predicted label. Some misclassification pairs might be more problematic than others. We quantify the extent these misclassification which could be problematic, i.e., class level bias. Let us suppose we quantify class level bias for a true class  $c_{true}$  predicted to be label  $c_{predicted}$  with the following notation:

$$\beta(c_{true} \xrightarrow{b} c_{predicted})$$

To quantify class level bias, we devise a function for  $\beta(c_{true} \xrightarrow{b} c_{predicted})$  for each misclassification pair. We investigated class-level bias of a BERT [10] model used for a chatbot’s intent classification. The main contribution of this research is a bias measure called *directional pairwise class confusion bias* that evaluates the extent of a trained model’s bias between a pair of classes, favoring one against the other. We also visualize

our bias using heatmaps. Color densities were used to recognize the most critical biased pairs. Directional pairwise class confusion bias is an aggregate effect of class imbalance in the training data or semantic biases encapsulated through accents, grammatical mistakes, misspelling or chatbot’s limited domain knowledge.

## 1.4.2 Quantify Unfavorable Stereotype Bias

Societal stereotype biases [37] are preconceived opinions, attitudes and judgement that are either positive or negative attributes to a group of people identified by various demographics like gender, age, race, ethnicity, language, nationality, disability, sexual orientation, etc. Unfavorable type of stereotype bias are not always based on facts and can be relentlessly abusive. While we would not want the NLP model to learn unfavorable societal stereotype biases, the model should be able to learn categorical biases that signify general knowledge of the world. For example, it is important that the model embed ‘United States of America’ and ‘Canada’ similarly because geographically these two countries border each other. However, we do not want the model to embed ‘United States of America’ closely to White people but farther to people of other races.

We show via experiments on the last 4 layers of BERT (*bert-base-uncased*) that most popular bias evaluation measures like WEAT [35] and SAME [36] cannot quantify categorical bias (general knowledge and facts) effectively. We also find inconsistencies for their results when measuring unfavorable stereotype bias. We introduce *Differential Cosine Bias (DiCoBi)* and show via experiments that it can appropriately quantify categorical bias and stereotype bias. DiCoBi measure utilizes four sets of words: anchor sets ( $\mathbf{X}, \mathbf{Y}$ ) and **test sets** ( $\mathbf{A}, \mathbf{B}$ ). Our DiCoBi measure evaluates whether the difference between  $\mathbf{A}$  and  $\mathbf{B}$  is similar to the difference between  $\mathbf{X}$  and  $\mathbf{Y}$ . For example, if  $\mathbf{A}$  and  $\mathbf{B}$  indicate different genders, the similarity of their difference to the difference of  $\mathbf{X}$  and  $\mathbf{Y}$  would indicate gender bias in professions if  $\mathbf{X}$  and  $\mathbf{Y}$  represent professions.

### 1.4.3 Quantify Domain Knowledge

Most of the research works in NLP bias have solely focused on measuring and mitigating human stereotype biases in NLP systems [38, 35, 39, 40, 34, 41, 42, 32]. We define domain bias as the ability of a language model to disambiguate the correct domain meaning for polysemic words having a very different meaning in the layman sense. Especially when some NLP models are built only for a specific domain, it is essential that the domain biases are present. It should be noted that societal stereotype biases are not desired in NLP models and need to be mitigated, but domain biases are vital for the purpose of the NLP model application.

Domain bias can be evaluated i) using a machine learning task whether the performance of the system improves as the model is trained to learn the domain or ii) checking the tendency of polysemic words to their domain interpretations. In this dissertation, we analyze the tendency of the word embeddings. For example, if we are building a language model for an insurance domain, the word *Umbrella* should mean the type of insurance coverage instead of its layman meaning, an accessory that humans use to protect themselves from rain or heat (Figure 1.1). After training the language model on insurance text corpora, we want to quantify by how much the domain meaning of *Umbrella* is being interpreted by the model. In other words, we quantify to what extent the model is being customized for its domain application. Our method uses three sets of words:  $x$  as the polysemic word,  $A$  as the related words to the layman meaning, and  $B$  as the domain relevant words. Hence, we propose a novel measure *domain\_gain* for quantifying domain knowledge whether  $x$  is closer to  $A$  or  $B$ .

## 1.5 Copyrights and Permissions

A substantial amount of text, figures and tables included in this dissertation are from papers whose primary author is the same as this dissertation, and published by IEEE and World Scientific. The following papers are reprinted with permission:

- © 2022 IEEE. Reprinted, with permission, from Sayenju, Sudhashree, Ramazan

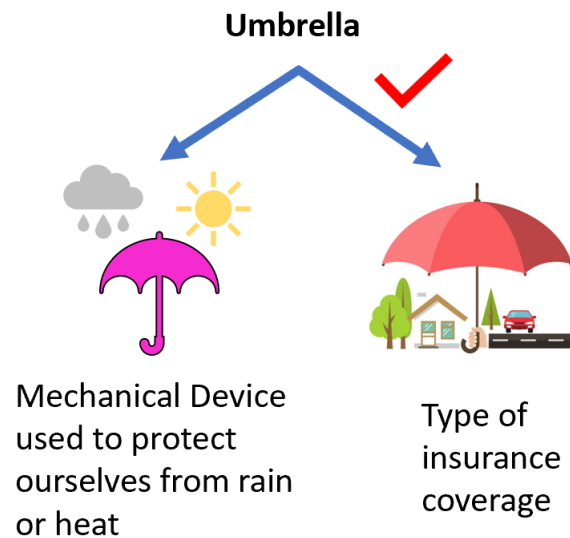


Figure 1.1: Within an insurance domain, we want *Umbrella* to mean the type of insurance coverage instead of device used to protect from rain or heat

Aygun, Jonathan Boardman, Duleep Prasanna Rathgamage Don, Yifan Zhang, Bill Franks, Sereres Johnston, George Lee, Dan Sullivan, and Girish Modgil. "Directional pairwise class confusion bias and its mitigation." In 2022 IEEE 16th International Conference on Semantic Computing (ICSC).

- *Quantification and Mitigation of Directional Pairwise Class Confusion Bias in a Chatbot Intent Classification Model* by Sayenju, Sudhashree, Ramazan Aygun, Jonathan Boardman, Duleep Prasanna Rathgamage Don, Yifan Zhang, Bill Franks, Sereres Johnston, George Lee, Dan Sullivan, and Girish Modgil. In International Journal of Semantic Computing: Vol. 16, No. 04, pp. 497-520. © 2022 World Scientific.
- © 2022 IEEE. Sayenju, Sudhashree, Ramazan Aygun, Bill Franks, Sereres Johnston, George Lee, and Girish Modgil. "Stereotype and Categorical Bias Evaluation via Differential Cosine Bias Measure." In 2022 IEEE International Conference on Big Data (Big Data).
- © 2023 IEEE. Sayenju, Sudhashree, Ramazan Aygun, Bill Franks, Sereres Johnston, George Lee, and Girish Modgil. "Quantifying Domain Knowledge in Large Language Models." In 2023 IEEE Conference on Artificial Intelligence (CAI)

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Kennesaw State University's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

## 1.6 Dissertation Organization

This dissertation is organized as follows:

- Chapter 2 presents literature review of research that has been done in the field of NLP bias detection and mitigation. We also present some limitations of existing bias evaluation methods.
- In Chapter 3 we describe our novel technique for detecting class level bias called *directional pairwise class confusion bias*. Additionally, we also present two mitigation strategies to reduce this bias to show how this bias measure could be utilized.
- Chapter 4 introduces *Differential Cosine Bias (DiCoBi)* measure for detecting stereotype and categorical bias (general knowledge and facts). Experiments are provided to show the DiCoBi measure works for quantifying categorical as well as stereotype bias.
- Chapter 5 elaborates how domain knowledge is quantified and compared across multiple models. We describe our *domain\_bias* measure to quantify whether a model's tendency lie more towards the layman or domain on specific chosen words.
- Chapter 6 summarizes the findings of this research and potential future work.



# Chapter 2

## Related Work

This chapter provides an overview related work conducted on different types of biases in Large Language Models (LLMs). We first begin by describing research works that recognize various types of bias present in LLMs. Secondly, we present research work related to class level bias followed by summary of stereotype bias studies in the most popular LLMs like Word2Vec [43, 44], GloVe [45], CBOW [44], ELMo [46], BERT [10] and GPT [11]. After presenting some widely used bias evaluation measures and their limitations, we describe research work that quantifies model tendency after customization.

### 2.1 Types of Biases

There are many NLP applications that directly affect humans. Thus, before deploying LLMs in real world applications, it is essential to evaluate bias in those models and when possible even mitigate the unfavorable kinds of bias it. However, the first step in the bias analysis process needs to be awareness about what bias means. Cognitive biases are the flaws in human thinking that deviate from logic or rationality [37].

Bias can be introduced at several points in the machine learning pipeline, and Suresh et al. [47] provide a useful taxonomy of the corresponding biases. Shah et al. [48] mention four situations in the supervised NLP pipeline, specifically where bias can occur. They can be listed as label bias, selection bias, representation bias, and over-amplification. Label bias occurs in annotating training labels. Selection bias takes place in sampling ob-

servations. Representation bias occurs when a model incorrectly compares two situations. Finally, over-amplification is considered a bias that is associated with the machine learning hypothesis. Dixon et al. [49] introduce a method to measure and mitigate unintended bias in text classification models. They contrast unintended bias with fairness which is a measure of potentially negative impact on society. According to Dixon et al. [49], unintended bias is caused by the disproportional representation of demographic identity terms in training data.

One of the biggest challenges in bias studies is the lack of an adequate bias definition in research studies leading to vague interpretation of experimental results. Blodgett et al. [50] surveyed 146 papers that evaluate bias in NLP. They also state that most of the papers do not provide a clear definition of bias they are interested in. Thus, they argue this as the reason why the quantitative techniques (both for bias detection and mitigation) used in many papers poorly matched their motivation. Blodgett et al. [50] provide three recommendations. The first recommendation mentions the need to explore literature outside of NLP especially in sociology, social psychology, sociolinguistics, and linguistic anthropology to better understand the relationship between language and social hierarchies. Eventually this recommendation should enable us to understand and define bias more concretely. The second recommendation states to explicitly define why some NLP system traits are harmful and to whom specifically [50, 51]. The last recommendation suggests to learn about language used by community members who are affected by biased NLP systems. This last one points out the need to examine the relations between NLP researchers and affected communities.

There have been also attempts to contextualize bias and accomplish transparency on bias analysis. Muñoz et al. [52] narrow its focus on NLP literature that involves bias in LLMs only. Their work formalizes the definition of bias in different contexts. It distinguishes the stereotype bias, LLM type, training data, language of model, bias evaluation techniques, mitigation strategies, modelling stages and NLP task type of each paper and neatly summarizes them in six tables. Instead of providing generic suggestions as [50] that requires no particular order, Muñoz et al. [52] provide a step by step methodology to

handle bias in LLMs. This guide includes seven steps that starts from defining stereotyped knowledge and ends with reporting the entire procedure by attaching documents such as data sheets and model cards to achieve transparency.

Among all the different types of unfavorable stereotype biases, the most common bias that was assessed in the literature is gender bias. Sun et al. [53] survey various papers that detect and make efforts to mitigate gender bias. Like many other biases, gender bias has been found to exist in all stages of the NLP modelling such as training data, pre-trained models and algorithms [35, 54, 55, 56]. Sun et al. [53] state that gender debiasing methods in NLP focus only on one segment of the NLP pipeline. It is yet to be identified how different parts of the NLP pipeline can be used to mitigate gender bias overall. Since research on gender bias in NLP has merely been a budding field, gender bias has only been observed in a limited number of NLP applications [55, 57]. Sun et al. [53] also point out that gender bias analysis in NLP is limited to English language, binary gender (male and female) and non-interdisciplinary collaboration. Thus, this bias requires further study of identifying bias and development of mitigation techniques [53].

Race is another unfavorable stereotype bias that was evaluated in various research papers. Unfortunately, race in NLP is not explored as much in detail as gender. This is probably the result of even less interest in socio-economic status, disability and sexual orientation [58, 59]. Although race is a global construct, how people identify their race is very subjective and differs from country to country. Field et al. [60] surveyed 79 papers in the ACL anthology that identify or even make efforts to mitigate racial bias. Their work also points out that minority races are mostly excluded from NLP research both in the creating and consuming end of technology. Additionally, Field et al. [60] also summarize that NLP research on race is limited to a small set of sub-tasks and definitions of race which conceals harm and might give the impression of natural or normal. Lastly, similar to gender bias, racial bias has also been detected in different stages of NLP pipeline [35, 54, 61, 62, 63, 64, 65, 49].

Although both Blodgett et al. [50] and Muñoz et al. [52] are very comprehensive in their survey, they mostly analyze each NLP bias paper individually and extract shortcom-

ings and strengths from each paper. Therefore, both papers made suggestions based on the shortcomings found in majority of NLP bias research studies.

## 2.2 Class Level Bias

For any machine learning model that makes decisions involving humans, inspecting the model's bias and fairness becomes very crucial. Detecting as well as mitigating bias is important. AI Fairness 360 (AIF360) [25] is an open source Python toolkit that provides various bias metrics and algorithms to mitigate bias in structured datasets and models. AIF360 includes over 71 bias detection metrics and 9 bias mitigation algorithms. Additionally, it also includes a unique extensible metric explanations facility to help consumers of the system understand the meaning of bias detection results. Although AIF360 is a very comprehensive tool, its bias detection and mitigation only works for structured data that contain protected attributes. Alternatively, Amazon Web Services (AWS) clients can make use of Sagemaker's Clarify [26]. Clarify offers explainability, bias detection and bias mitigation. Clarify can schedule recurring jobs to monitor bias drifts and give explanations. The bias monitor includes 21 bias detection metrics and 4 bias mitigation algorithms. Although both AIF360 and AWS Sagemaker Clarify offer bias detection and bias mitigation techniques, their bias metrics and mitigation algorithms are designed for protected attributes included in the features dataset. These tools are efficient and easy to measure unfavorable stereotype bias in presence of protected class. In such cases, the ground truth for unfavorable stereotype bias is known. However, they do not highlight class level bias for the trained model.

While there is research that solely addresses the class imbalance problem [27, 28, 29, 30], not much research has been conducted in model's classification bias due to various anomalies enclosed in the text semantics. A model could favor a certain class more than another in a task like intent classification. Therefore, defining and quantifying such class-specific bias can help find the cause of the bias and eventually find procedures to mitigate it.

## 2.3 Unfavorable Stereotype Bias in LLMs

There are numerous cognitive biases that can be identified based on domains like social, behavioral and more. Stereotyping is type of a cognitive bias when assumptions are made or discrimination takes place on the basis of national, ethnic or gender groups [37]. On the other hand, model bias looks for whether a model has preference for certain classes or data groups.

Although there are a large number of types of cognitive biases, this section focuses on unfavorable stereotype bias. Stereotype bias is generally regarded as unfavorable since it causes discrimination. Often implicit bias leaves the minority group at disadvantage although the bias is not based on facts and is relentlessly abusive. Stereotyped bias attributes are either positive or negative attributes to a group of people identified by various demographics like gender, age, race, ethnicity, language, nationality, disability, sexual orientation, etc. While stereotype bias based on attribute such as age could be unfavorable for job hiring, it could be favorable for auto insurance as teenager novice drivers are more likely to have accidents and pay higher premiums than safe drivers. The domain experts should decide whether such bias is favorable or not.

When collecting structured data from humans, stereotype bias is present in protected attributes such as age, gender, race, ethnicity, religion, profession, etc. The performance of a model can vary for different values of the same protected attribute. For example, if the protected attribute was gender, the performance of a model for male instances might be better than for female instances.

We survey bias analysis from the perspective of various LLMs. We include the most widely used LLMs of the past few years such as Word2vec [43, 44], GloVe [45], CBOW [44], ELMo [46], BERT [10] and GPT-2 [11]. Popular word embeddings like Word2Vec [66] and Glove [45] have found to inherit unfavorable gender, race and religion bias from the corpus they were trained on [17, 67, 68, 69]. Apart from word embeddings, language models like BERT [70, 71, 72, 73] and GPT-3 [74] have been found to have unfavorable stereotype bias too. Although many variants of these models exist, we will be focusing on research studies that include the original models at some stage in

their analysis. The unfavorable stereotype bias analysis for each model is summarized in a table and as a graph. In the tables, unavailable information is provided as NA (Not Applicable). The graphs enable to reduce information duplicity of the table and give a clear and condensed overview of various issues and associations of each model. In each graph the models are represented in yellow rectangles, attributes of the model bias research are depicted in pink ovals, and values of the attributes are shown in blue ovals.

In this section, at first, we summarize unfavorable stereotype bias analysis from the perspective of each LLM. Secondly, we show how analysis of bias has evolved over time. Finally, we present analysis of various unfavorable stereotype biases associated with the most popular LLMs.

### **2.3.1 Word2vec**

Developed by Google and published over two papers in 2013, Word2vec [43, 44] was a progressive approach in NLP to vectorize words semantically using a two-layer neural net. Word2vec is still the foundation of the most modern advances in NLP such as ELMo [46] and BERT [10]. The original Word2vec model was mostly trained on Wikipedia corpus [75]. Thus, it includes grammatically correct sentences, no spelling errors, and systemic bias from historic texts. Word2vec has also been adapted in a wide variety of applications by performing transfer learning (using model trained for a specific task in a different but related task). Keeping these factors in mind, researchers suspected various kinds of biases would inherit in Word2vec from the data itself. Moreover, there is also a risk that the complex algorithms used in deep neural networks may display biases that were not present in the data.

Table 2.1 lists a few papers that evaluated bias in Word2vec models by training or testing them on various datasets. Although different types of biases were evaluated in each, mitigation techniques were not proposed in all papers. These research studies mostly aimed at gender and ethnicity bias. The primary language of bias analysis was English except the work conducted by Díaz et al. [78] (Spanish). For mitigation strategies on Word2vec, we only see vector space manipulation used as a mitigation mechanism.

LLM	Reference	Data	Stereotype	Language	Evaluation	Mitigation
Word2vec	Bolukbasi et al. [54]	GoogleNews corpus (w2vNEWS), Common Crawl	Gender	English	Analogies, Cosine Similarity	Vector Space Manipulation
	Manzini et al. [76]	Reddit L2 corpus	Ethnicity, Gender, Religion	English	PCA, WEAT, MAC, Clustering	Vector Space Manipulation
	Caliskan et al. [35]	Common Crawl, Google News Corpus, Occupation Data (BLS)	Ethnicity, Gender	English	Association Tests (WEAT, WEFAT)	NA
	Swinger et al. [77]	Google News, Web data, First Names (SSA)	General	English	WEAT	NA
	Díaz et al. [78]	Wikipedia-es 2006	Gender	Spanish	Analogies	NA
	Cheng et al. [79]	Perspective API's Jigsawdataset	Gender	English	PCA	Vector Space Manipulation, (Toxicity debias)
	Curto et al. [80]	Google News	Gender	English	Analogies, Cosine Similarity	NA
	Chen et al. [81]	Wikipedia, Book Corpus, GLUE	Gender, Profession	English, Mandarin Chinese, Spanish, English, Arabic, German, French, Farsi, Urdu and Wolof	Analogies, Average Cosine Similarity	NA
	Chen et al. [82]	Chinese Wikipedia Dump, CSemBias Dataset.	Gender	Chinese	Accuracy, Analogies, Word similarity	Vector Space Manipulation,

Table 2.1: Papers analyzing bias in Word2vec.

In order to remove repeated information in the table and get a clearer view of bias analysis research done in Word2vec, we have created a visual graph (Fig. 2.1) that summarizes all the properties deduced from the bias research on Word2vec. Fig. 2.1 shows that bias evaluation research in Word2vec is limited to only English and Spanish languages. Similarly, bias mitigation techniques are limited to vector space manipulation. The bias evaluation techniques range from simple techniques like cosine similarity and Principle Component Analysis (PCA) to more sophisticated ones such as Word Embedding Association Test (WEAT) [35]. The unfavorable stereotype biases under scrutiny in the research studies included specific ones like gender, ethnicity, religion and also general bias. Due to the fact that Word2vec was developed in 2013 and its popularity has been surpassed in the years following it by more complex language models like transformers and BERT, we do not see a large variety in its bias research.

### 2.3.2 GloVe

Other than Word2vec, another popular word embedding model is GloVe [45]. GloVe (short for Global Vector) uses an unsupervised machine learning algorithm to generate its word embeddings. Developed by the Computer Science department in Stanford Uni-

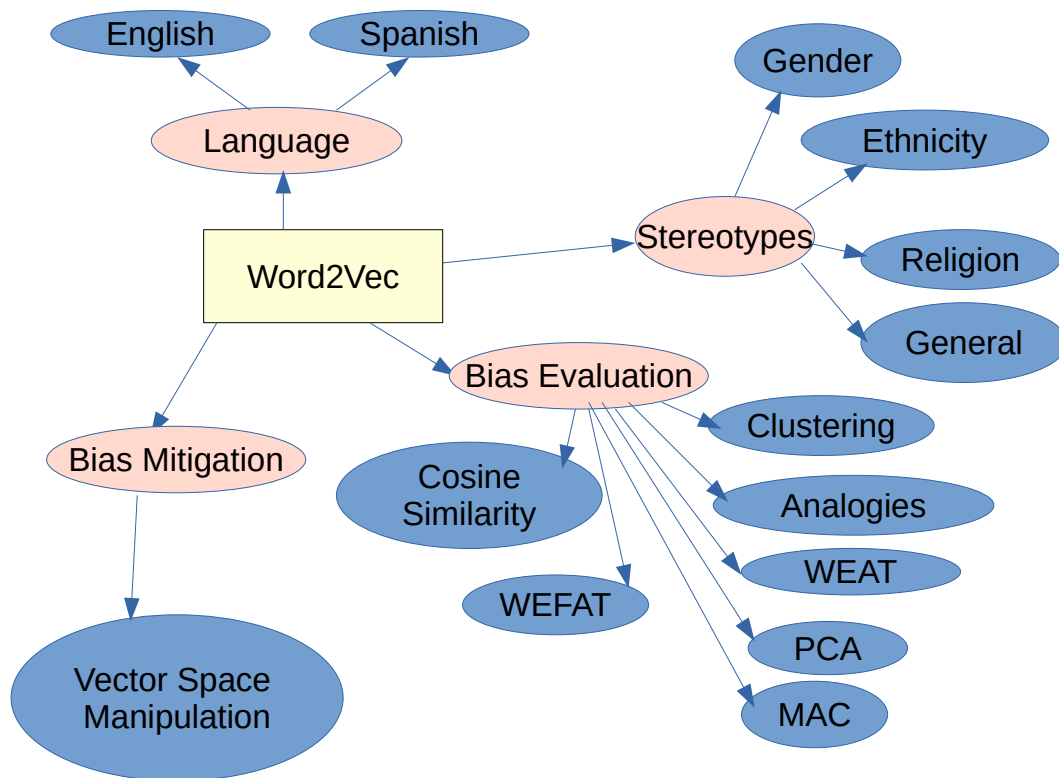


Figure 2.1: Summary of bias analyses done in Word2Vec based on published papers.

versity, GloVe is also trained on multiple text corpora like Wikipedia, Common Crawl and Twitter. Similar to the case of Word2vec [43, 44], this data-based learning approach of GloVe also makes it susceptible to bias inherited from the training data. Additionally, the global word-word co-occurrence statistics GloVe implements might account for algorithmic bias.

More papers have been published on bias analysis of GloVe than for Word2vec. Table 2.2 enlists the publications on bias in GloVe. Larger number of research studies also account for larger variety in the number of datasets, languages, bias evaluation techniques as well as mitigation techniques. Common Crawl and Wikipedia were the most frequent datasets used in published works. Gender bias was the most common stereotype bias present in GloVe. Unlike in Word2vec bias research which included language models in English and Spanish only, five more languages (German, Italian, Russian, Croatian, Turkish) were covered for GloVe. Prediction accuracy and WEAT were the most common



LLM	Reference	Data	Stereotype	Language	Evaluation	Mitigation
GloVe	Zhao et al. [56]	OntoNotes 5.0, WinoBias, Occupation Data (BLS), B&L	Gender	English	Prediction Accuracy	Data Augmentation (Gender Swapping), Vector Space Manipulation
	Zhao et al. [83]	2017 English Wikipedia dump, SemBias	Gender	English	Prediction Accuracy, Analogies	Attribute Protection, Vector Space Manipulation, Hard-Debias
	Escudé et al. [84]	United Nations [92], Europarl [93], newstest2012, newstest2013, Occupation data (BLS)	Gender	English, Spanish	BLEU	Vector Space Manipulation (Hard-Debias)
	Lauscher et al. [85]	NA	General	German, Spanish, Italian, Russian, Croatian, Turkish, English	WEAT, XWEAT, ECT, BAT, Clustering (KMeans) (BIAS ANALOGY TEST)	Vector Space Manipulation, DEBIE
	Dev et al. [86]	Common Crawl	Gender	English	WEAT*, SIRT	Vector Space Manipulation, OSCaR
	Caliskan et al. [35]	Common Crawl, Google News Corpus, Occupation Data (BLS)	Ethnicity, Gender	English	Association Tests (WEAT, WEFAT)	NA
	Swinger et al. [77]	Google News, Web data, First Names (SSA)	General	English	WEAT	NA
	Dev et al. [87]	Wikipedia Dump, WSim-353, SimLex-999, Google Analogy Dataset	Gender, Age, Ethnicity	English	WEAT, EQT, ECT	Vector Space Manipulation
	Lauscher et al. [88]	English Wikipedia, Common Crawl, Wikipedia, Tweets	Gender	English, German, Spanish, Italian, Russian, Croatian, Turkish	WEAT, XWEAT	NA
	Guo et al. [89]	CommonCrawl, Billion Word Benchmark, BookCorpus, English Wikipedia dumps, BookCorpus, WebText, Bert-small-cased	Intersectional Bias (Gender, Ethnicity)	English	WEAT, CEAT	NA
	Zhao et al. [90]	One Billion Word Benchmark, WinoBias, OntoNotes 5.0	Gender	English	PCA, Prediction Accuracy	Data Augmentation, Attribute Protection (gender swapping averaging)
	Curto et al. [80]	Wikipedia, Twitter	Gender	English	Analogies, Cosine Similarity	NA
	Marcé et al. [91]	Twitter	Gender	English	Confusion matrix, F1-score	NA

Table 2.2: Papers analyzing bias in GloVe.

bias evaluation techniques used. Additionally, more than half of the papers also offer different types of bias mitigation strategies.

Fig. 2.2 shows the visual summary of Table 2.2 as a graph. In comparison to the graph of Word2vec (Fig 2.1), it is evident that there are more blue ovals (bias research attributes). Less commonly spoken languages like Croatian, Turkish, Russian, German and Italian were also used for building and testing the GloVe embeddings. There were a wide variety of evaluation techniques to quantify various types of stereotype bias. Among the 14 bias evaluation techniques, many of them were modern sophisticated metrics such as WEAT [35], WEFAT [35], BAT [85], SIRT [86], ECT [87], EQT [87], CEAT [89] and BLEU [94]. Other than vector space manipulation, the bias mitigation strategies include data augmentation, attribute protection and hard-debias.

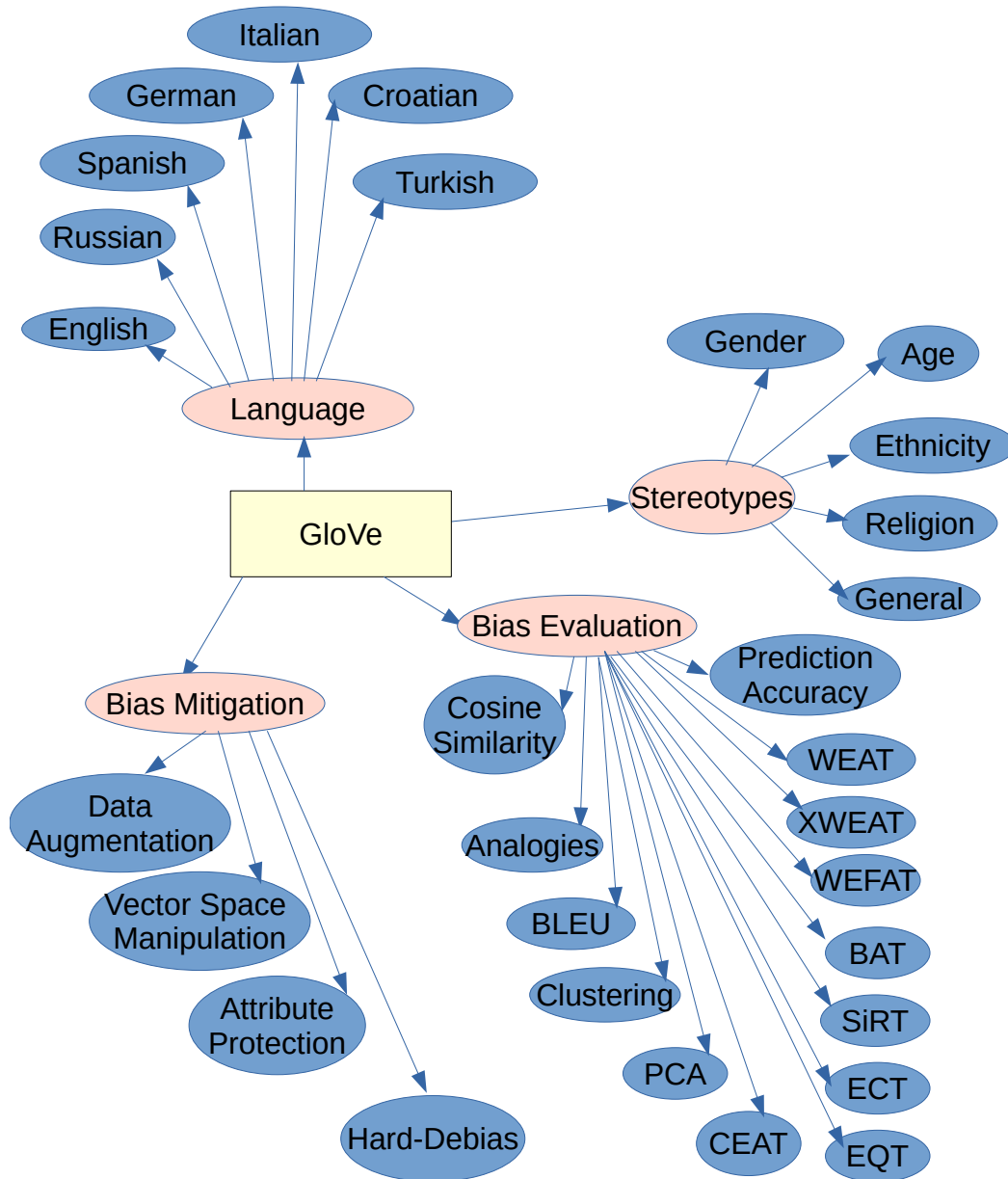


Figure 2.2: Summary of bias analyses done in GloVe based on published papers.

Although Word2vec and GloVe do not differ conceptually to a large extent, the variety in research attributes in GloVe shows there was higher level of interest in researching about bias for GloVe than Word2vec. The added benefit of GloVe is that it is easier to parallelize the code for extremely large datasets. Thus, this reason might have contributed to GloVe being more prevalent in the NLP research community compared to Word2vec.

### 2.3.3 CBOW

Continuous Bag of Words (CBOW) model is the architecture devised and used by Mikolov et al. [44] to train Word2vec. This architecture uses a number of words that appear before and after the word in the middle to predict the semantic word embedding of the middle word. This way CBOW was designed to encapsulate the contextual semantics of a word in its embedding. The final vocabulary of CBOW only gives one embedding for each word, although a word might have very different semantic meanings depending on context. Additionally, this architecture is very data driven which makes it prone to inheriting stereotype bias from the training data.

LLM	Reference	Data	Stereotype	Language	Evaluation	Mitigation
CBOW	Lausher et al.[85]	NA	General	German, Spanish, Italian, Russian, Croatian, Turkish, English	WEAT, XWEAT, ECT, BAT, Clustering (KMeans) (BIAS ANALOGY TEST)	Vector Space Manipulation, DEBIE
	Lausher et al.[88]	English Wikipedia, Common Crawl, Wikipedia, Tweets	Gender	German, Spanish, Italian, Russian, Croatian, Turkish, English	WEAT, XWEAT	NA
	Hall et al. [95]	English Gigaword, Wikipedia, Google Analogy, SimLex-999	Gender	English	Analogies, WEAT, Sentiment Classification, Clustering	Hard-Debiasing, CDA, CDS
	Lausher et al.[96]	translated WEAT test set, Leipzig news, Wikipedia, Twitter, CommonCrawl	Gender, Ethnicity	Modern Arabic, Egyptian Arabic	WEAT, XWEAT, AraWEAT, ECT, BAT	NA
	Leavy et al.[39]	British Library Digital corpus, The Guardian article	Gender	English	Association, Prediction likelihood, Sentiment Analysis	NA
	Marcé et al.[91]	Twitter	Gender	English	Confusion matrix, F1-score	NA
	Chen et al.[82]	Chinese Wikipedia Dump, CSemBias Dataset.	Gender	English	Accuracy, Analogies, Word similarity	Vector Space Manipulation

Table 2.3: Papers analyzing bias in CBOW.

Table 2.3 enlists the NLP bias papers that used CBOW to get word embeddings in various languages. Here again, gender was the most common stereotype bias being quantified. Similarly, text corpus from Wikipedia and CommonCrawl appeared multiple times in the literature. Since, CBOW is an architecture and not a pre-trained embedding model, it is more accessible to be used for other languages instead of being limited to English. Most papers evaluated stereotype bias using WEAT [35] and its cross lingual extension XWEAT [85]. Not every research that evaluated NLP bias using a CBOW model pre-

sented bias mitigation techniques of their own.

At first glance, the bias literature summary graph of CBOW (Fig. 2.3) also displays a lot of blue ovals (bias research attributes) similar to that of GloVe (Fig. 2.2). Thus, the bias research explores different dimensions of research. In addition to English, less explored languages like German, Spanish, Italian, Russian, Croatian, Turkish as well as versions of Arabic namely Modern and Egyptian Arabic were explored. Only three types of bias stereotype were analyzed for CBOW namely gender, general and ethnicity. The bias evaluation methods range from simple techniques like clustering and analogies to more intricate techniques like WEAT, XWEAT, AraWEAT and ECT. Four bias mitigation techniques used for CBOW were vector space manipulation, hard-debias, CDA (Contextual Data Augmentation) [97] and CDS (Contextual Data Substitution) [95].

### 2.3.4 ELMo

Embeddings from Language Models (ELMo) [46] is an NLP framework that developed a new variation of the Long Short-Term Memory (LSTM) [98] architecture. ELMo was developed by AllenNLP [99] team of Allen Institute for AI (AI2) in 2018. ELMo's novelty is that the embedding of a word varies for different contexts. The context of each word is determined from the words appearing before it as well as after it. ELMo does this by concatenating the forward language model layers with the respective backward language model layers. ELMo also uses character level tokens to generate word embeddings to feed the main model. This enables ELMo to be easily adapted for various NLP tasks. Training ELMo for multiple NLP capabilities required training on a large text corpus like 1B Word Benchmark [100]. With the large architecture and training corpus in use, ELMo is also prone to storing stereotype bias information.

The publications on stereotypical bias found in ELMo are gathered in Table 2.4. Unlike the bias research done on GloVe or CBOW, the publications on ELMo are not versatile. Out of the five papers two (May et al. [33], Tan et al. [34]) do not reveal their data sources. Like in most NLP bias papers, the datasets CommonCrawl and Wikipedia were used for ELMo too. Zhao et al. [90] use the same 1B Word Benchmark [100] cor-

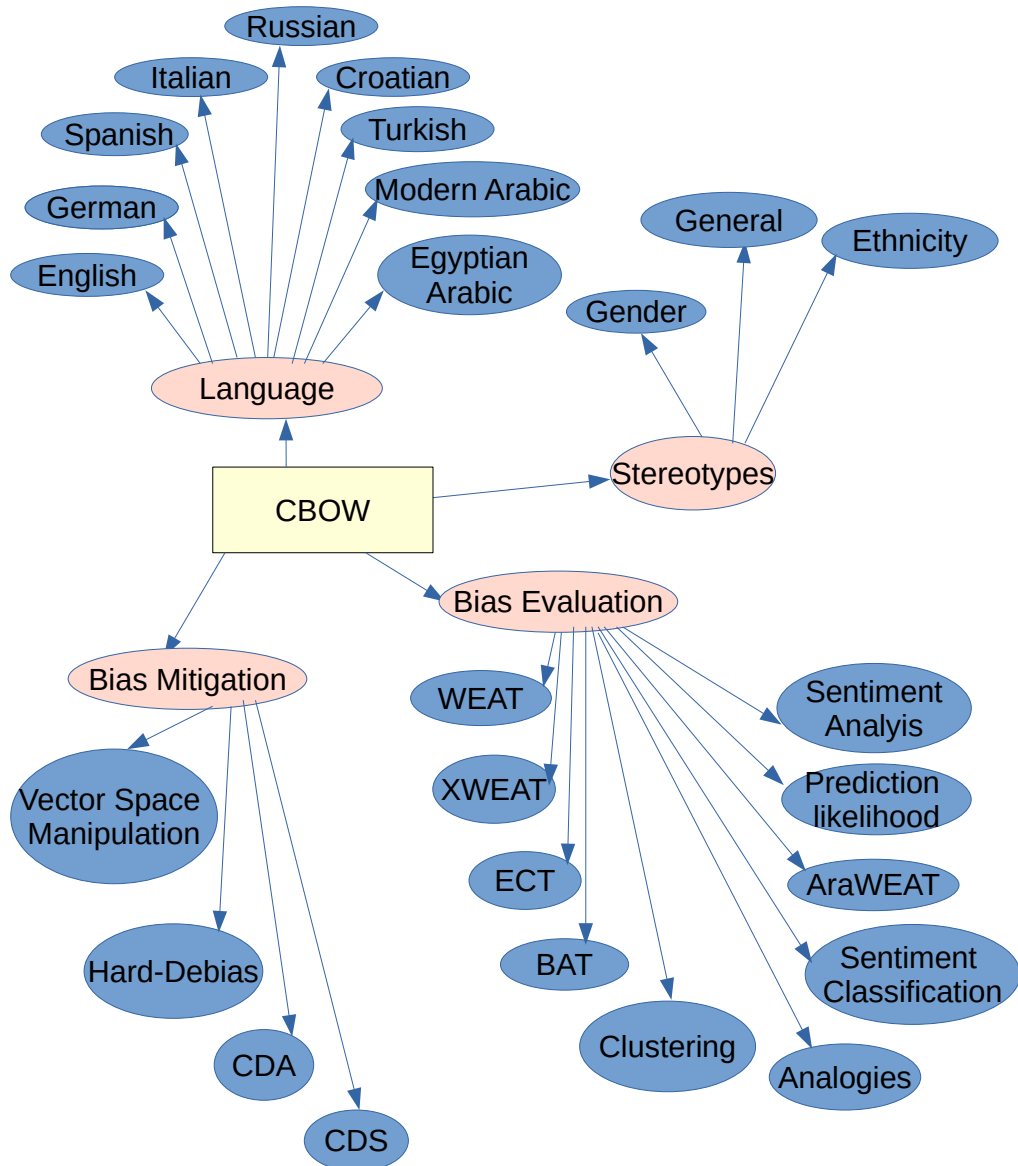


Figure 2.3: Summary of bias analyses done in CBOW through based on published papers.

pus ELMo originally used. Additionally, they use OntoNotes 5.0 and their own corpus Winobias. English is the only language used for modelling. The stereotype biases being analyzed were dominated by gender bias. Only Zhao et al. [90] proposed mitigation strategies. The evaluation methods ranged from simple methods like cosine similarity, PCA and prediction accuracy to more sophisticated metrics like WEAT, SEAT, CEAT.

Fig 2.4 provides the summary graph of bias research done for ELMo. We can immediately observe lower number of blue ovals in this graph especially compared to that of GloVe (Fig. 2.2) and CBOW (Fig. 2.3). In terms of languages, there is no diversity at all

LLM	Reference	Data	Stereotype	Language	Evaluation	Mitigation
ELMo	May et al. [33]	NA	Gender, Ethnicity	English	SEAT	NA
	Tan et al. [34]	NA	Gender, Race	English	Contextual SEAT	NA
	Guo et al. [89]	CommonCrawl, Billion Word Benchmark, BookCorpus, English Wikipedia dumps, BookCorpus, WebText, Bert-small-cased	Intersectional Bias (Gender, Ethnicity)	English	WEAT, CEAT	NA
	Zhao et al. [90]	One Billion Word Benchmark, WinoBias, OntoNotes 5.0	Gender	English	PCA, Prediction Accuracy	Data Augmentation, Attribute Protection (gender swapping averaging)
	Basta et al. [101]	English-German news WTM18	Gender	English	cosine similarity, clustering, KNN	NA

Table 2.4: Papers analyzing bias in ELMo.

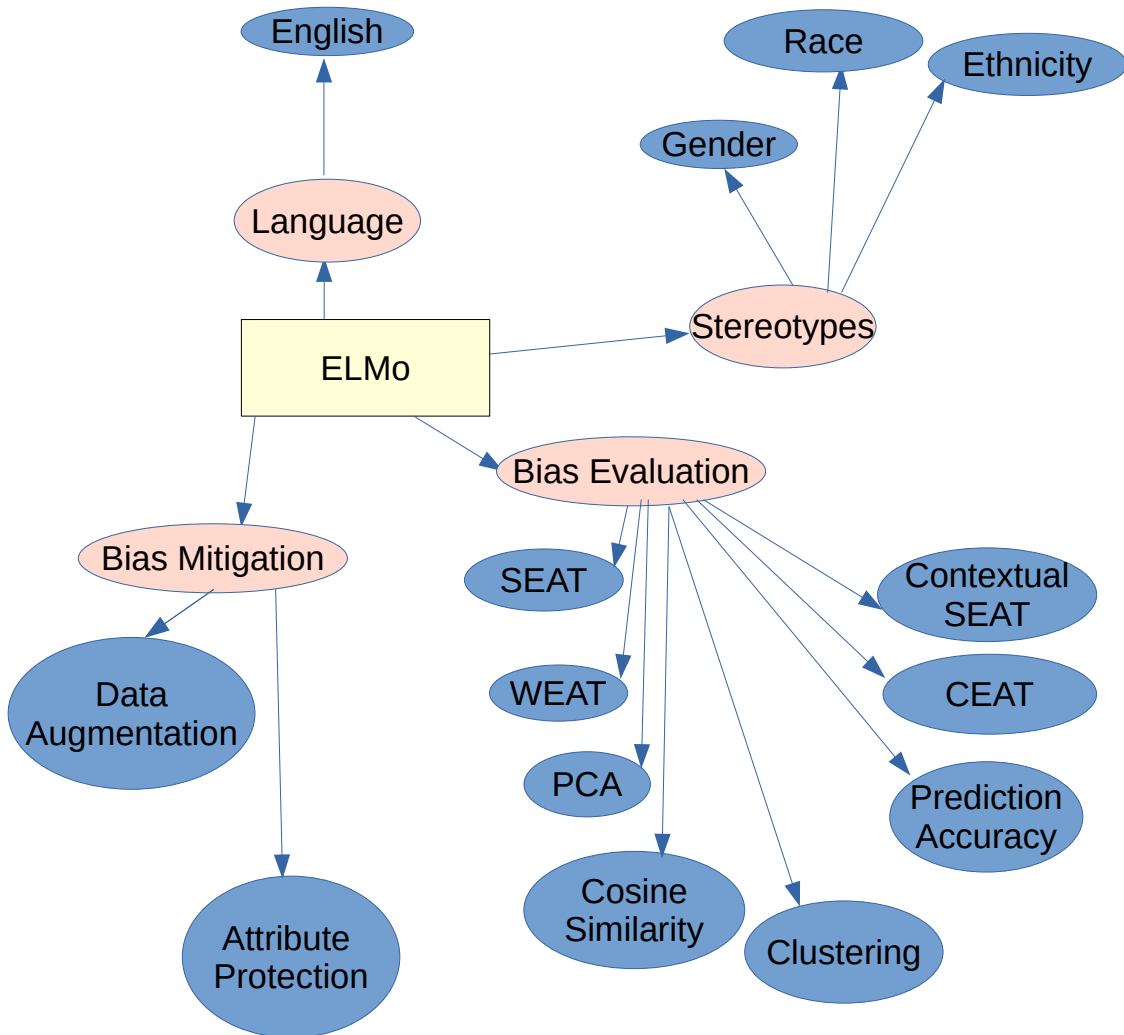


Figure 2.4: Summary of bias analyses done in ELMo based on published papers.

limited to only English language models. The stereotypes included the 3 most common types of biases, namely gender, race and ethnicity. Only two bias mitigation techniques were used on ELMo. There was a large assortment of bias evaluation techniques for ELMo.

The popularity of ELMo was soon outdone by Transformer and BERT. This is probably the reason why the bias research on ELMo was not explored in depth for various aspects. Although ELMo was very powerful in introducing the contextualized word embedding concept, the use of LSTM made it less compelling to use ELMo past the technologies that emerged after it.

### **2.3.5 BERT**

Bidirectional Encoder Representations from Transformers (BERT) [10] makes use of technology that existed before it called Transformer [9]. BERT was developed at Google AI language in 2018. Devlin et al. [10] show multiple experiments on various NLP tasks that outperformed the state of art models like OpenAI's GPT [102] and ELMo [46]. As the name suggests, the transformer nodes in BERT are bidirectional in all layers. This makes it more efficient than ELMo which concatenates the separate LSTM nodes for forward and backward context learning. On the other hand, although very similar, OpenAI's GPT was using only left to right context learning Transformer nodes. Thus, the dense architecture of BERT enables it to boost its learning capabilities. The corpus used in the pre-training phase of BERT were BooksCorpus and English Wikipedia. Both of these datasets contain a large amount of historical data with various potential biases. The complex architecture of BERT might also amplify some biases not present in the data. Fine-tuning BERT for various tasks used task specific text corpora. Those datasets add on their own types of biases. Although BERT showed striking results on standard NLP tasks, it lacked examples that highlight various types of unfavorable stereotype biases. The fame of BERT also gathered the interest of many NLP researchers who suspected of likely biases. Being aware of the biases in BERT is very important since it is used in numerous real world systems which are already causing harm to the society or might do

so in the future.

LLM	Reference	Data	Stereotype	Language	Evaluation	Mitigation
BERT	Vig et al. [31]	NA	Gender	English	Visualization, Text Generation likelihood	NA
	Bhardwaj et al. [32]	Equity Evaluation Corpus, Gen-data	Gender	English	EEC, Gender Separability, Emotion/Sentiment Scoring	Vector Space Manipulation
	May et al. [33]	NA	Gender, Ethnicity	English	SEAT	NA
	Tan et al. [34]	NA	Gender, Race	English	Contextualized SEAT	NA
	Nadeem et al. [39]	StereoSet	Gender, Profession, Race, Religion	English	CAT (Context Association Test)	NA
	Bartl et al. [103]	GAP, BEC-Pro, Occupation Data (BLS)	Gender	English, German	WEAT	Fine-tuning, CDS
	Sheng et al. [104]	One Billion Word Benchmark	Race, Gender, Sexual Orientation	English	Sentiment Score (VADER), Classification accuracy	Train LSTM/BERT
	Babacianjelodar et al. [105]	Wikipedia, Book corpus, Jigsaw identity toxic dataset, RtGender, GLUE	Gender, Race, Religion, Disability	English	Cosine Similarity, Accuracy, GLUE	Fine tuning
	Hutchinson et al. [59]	Jigsaw Unintended Bias	Disability	English	Sentiment Score	NA
	Kurita et al. [106]	Gendered Pronoun Resolution task, GAP Employee Salary Dataset for Montgomery County of Maryland	Gender, Profession	English	Log Probability Bias Score	NA
	Marcé et al. [91]	Wikipedia, Book corpus	Gender	English	Confusion matrix, F1-score	NA
	Herold et al. [107]	Wikipedia, Book corpus, WordNet	Disability	English	Log Probability Bias Score	NA
	Jentsch et al. [108]	Internet Movie Database	Gender	English	Absolute sentiment bias	NA
	Felkner et al. [109]	QueerNews	Sexual Orientation	English	log-likelihood, F1-score, Gender ratio	NA

Table 2.5: Papers analyzing bias in BERT.

Table 2.5 provides the literature that studies bias in BERT models. Out of the fourteen papers, three did not reveal their data sources. Except Wikipedia, other datasets used in the analyses are not commonly used. Analogous to papers involving other LLMs, literature on BERT’s bias also mostly focuses on gender bias. Less explored unfavorable stereotype biases like profession, disability and sexual orientation were also examined for BERT. English BERT models were used in all cases. German language BERT was included only in Bartl et al. [103]. Research studies used different methods to measure bias. The spectrum of bias evaluation methods include simple ones like accuracy and cosine similarity, as well as more modern techniques such as WEAT, SEAT and GLUE. Half of the work on BERT did not provide approaches for mitigation.

Considering the fact that BERT was published soon after ELMo and garnered more



fame than ELMo, it becomes obvious why the summary graph of BERT (Fig. 2.5) has more features (blue ovals) than for ELMo (Fig. 2.4). Although BERT was first published in 2018, it remains to be one of the most popular models till date. In terms of the languages BERT were modelled in, there was not much diversity. The studies only included German beyond English. Amongst all the models that preceded it, BERT showed the widest assortment of unfavorable stereotype biases being analyzed. It was also rich in the bias evaluation methods and mitigation techniques it used. Although, from Table 2.5 we know that half of them did not attempt to mitigate the bias, there still were four different mitigation procedures tried on BERT. All these methods can be applied either during training or post-training phase of modelling.

### **2.3.6 GPT Models**

Generative pre-trained transformers (GPT [102]) are LLMs developed by OpenAI. Till date there are four versions released namely GPT [102], GPT-2 [11], GPT-3 [12], GPT-4 [13]. OpenAI also released a user interface integrated version with chat functionalities called ChatGPT [110]. ChatGPT uses GPT-4.

#### **GPT**

In 2018, OpenAI released its first LLM called GPT [102]. This model has over 120 million trainable parameters and was trained on BookCorpus. GPT was released only a few months before its subsequent and up-scaled version GPT-2 [11] was released. Therefore, most of bias research has been dedicated to GPT-2. Table 2.6 summarizes the few papers that reviewed bias in GPT. Therefore, we do not observe variety in their research aspects. Only Guo et al. [89] listed its data source. Gender bias was observed in all papers. Other unfavorable stereotype biases analyzed included Ethnicity and Race. All models were based in English language. SEAT, Contextualized SEAT, WEAT, and CEAT methods were used or devised for unfavorable stereotype bias evaluation. There were no mitigation techniques used in any of the research works.

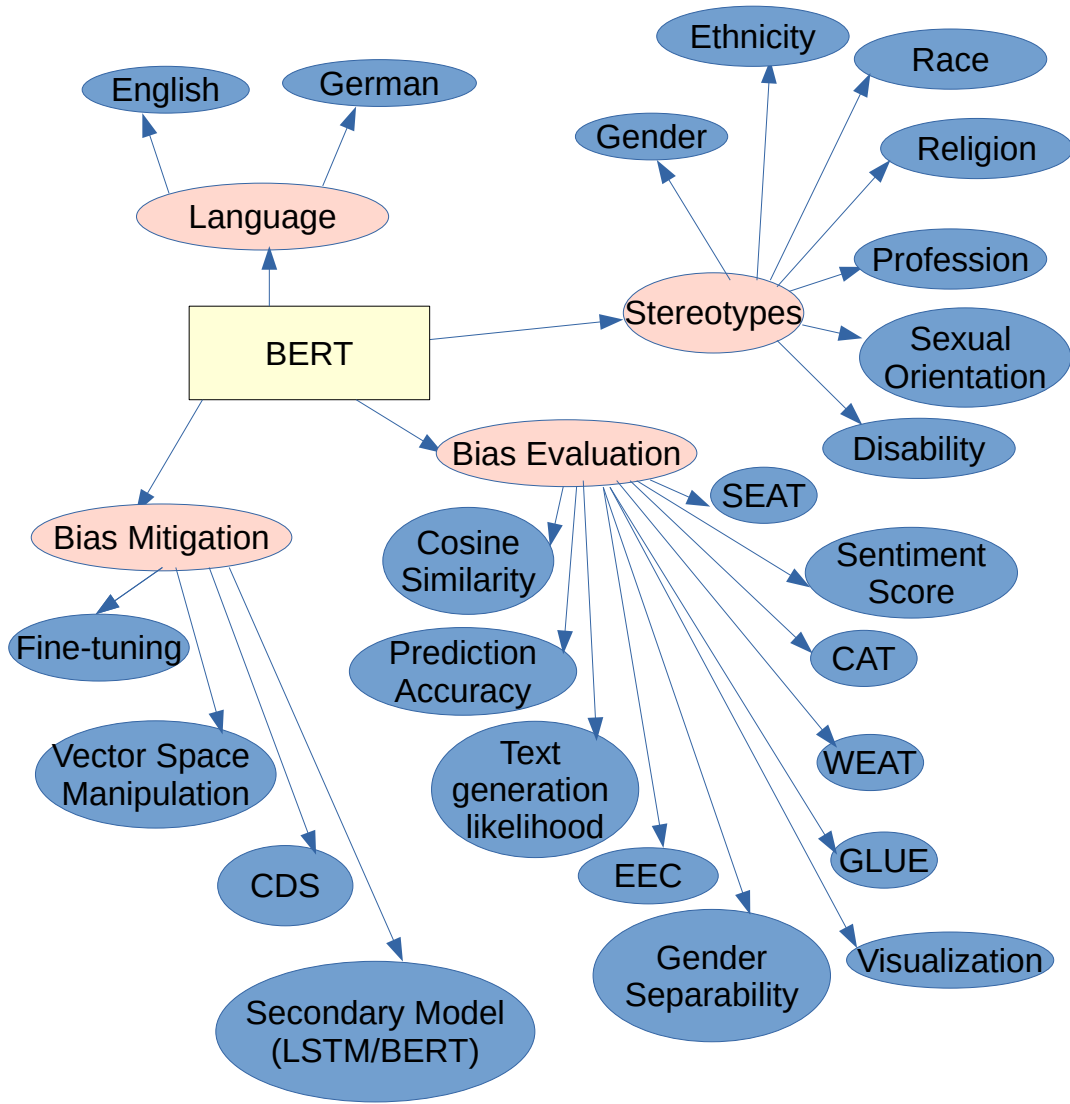


Figure 2.5: Summary of bias analyses done in BERT based on published papers.

LLM	Reference	Data	Stereotype	Language	Evaluation	Mitigation
GPT	May et al. [33]	NA	Gender, Ethnicity	English	SEAT	NA
	Tan et al. [34]	NA	Gender, Race	English	Contextual SEAT	NA
	Guo et al. [89]	CommonCrawl, Billion Word Benchmark, BookCorpus, English Wikipedia dumps, BookCorpus, WebText, Bert-small-cased	Intersectional Bias (Gender, Ethnicity)	English	WEAT, CEAT	NA

Table 2.6: Papers analyzing bias in GPT.

**GPT-2**

After the release of GPT [102] in 2018, its up-scaled version GPT-2 [11] was developed by OpenAI in 2019. GPT-2 has 10 times more learnable parameters and training data than the original GPT. Similar to BERT [10], GPT-2 also uses Transformer [9] nodes in its architecture. Their architecture has 1.5 billion learnable parameters and is capable of unsupervised learning. Unsupervised learning is performed via the language modelling where the next word is predicted based on given a set of starting words. Therefore, GPT-2 can be readily customized for various NLP tasks such as reading comprehension, translation, question answering and summarization. Their dataset is a 40GB Internet text containing 8 million web pages called WebText. The diversity of text domains also accounts for the diversity in the types of stereotype biases found in GPT-2. Despite the fact that the results presented in their paper [11] come close to humans, GPT-2 is not free from biases. Many research studies are dedicated to investigating various bias and a few even propose mitigation processes.

LLM	Reference	Data	Stereotype	Language	Evaluation	Mitigation
GPT-2	Vig et al. [31]	NA	Gender	English	Visualization, Text Generation likelihood	NA
	Tan et al. [34]	NA	Gender, Race	English	Contextualized SEAT	NA
	Nadeem et al. [39]	StereoSet	Gender, Profession, Race, Religion	English	CAT (Context Association Test)	NA
	Guo et al. [89]	CommonCrawl, Billion Word Benchmark, BookCorpus, English Wikipedia dumps, BookCorpus, WebText, Bert-small-cased	Intersectional Bias (Gender, Ethnicity)	English	WEAT, CEAT	NA
	Peng et al. [111]	Science fiction story corpus, Plotto, ROCstories, toxic and Sentiment datasets	Ethnicity	English	Classification Accuracy	Loss function modification
	Sheng et al. [104]	One Billion Word Benchmark	Race, Gender, Sexual Orientation	English	Sentiment Score (VADER), Classification accuracy	Train LSTM/BERT
	Groenwold et al. [112]	TwitterAAE, Amazon Mechanical Turk annotators (SAE)	Ethnicity	English	Text generation, BLEU, ROUGE, Sentiment Classification (VADER)	NA
	Honnavalli et al. [113]	Gender seniority compound bias dataset Amazon Mechanical Turk	Gender, Age	English	Percentage of Gendered search results	NA

Table 2.7: Papers analyzing bias in GPT-2.

Although the trained GPT-2 model on the entire dataset is not available for the gen-

eral public, a smaller version of GPT-2 was publicly released. The authors themselves had realized the potential misuses that could happen with GPT-2. Even with its smaller released version, GPT-2 grabbed the attention of NLP bias researchers. Table 2.7 provides a compilation of research studies that detected and in a few cases even mitigated the bias. Vig et al. [31] and Tan et al. [34] did not mention the datasets they used. A wide variety of very large text corpora were tested on GPT-2. Some popular ones include CommonCrawl, One billion word benchmark, Wikipedia and BookCorpus. In addition to gender, the most popular stereotype bias, GPT-2 also had profession, race, religion, ethnicity and sexual orientation biases. Guo et al. [89] explored the intersectional bias between gender and ethnicity. Even though GPT-2 is one of the advanced models, it was only used in English language. Popular bias evaluation methods like WEAT, sentiment score and classification accuracy were also utilized. Less common approaches to compute the bias were BLEU, ROUGE, CEAT, contextual CEAT, CAT and text generation likelihood. Only Peng et al. [111] and Sheng et al. [104] on GPT-2 attempted to mitigate the respective biases.

The summary graph of GPT-2's bias literature is presented in Fig. 2.6. In spite of eight papers reporting bias on GPT-2, the graph clearly shows the lack of variation in the languages GPT-2 was modeled on. Since GPT-2 was only released in English, NLP bias researchers were also limited to English language. Only two different mitigation techniques namely loss-function modification and secondary model were tried on GPT-2. However, in terms of types of unfavorable stereotype biases being analyzed and bias evaluation methods used, the research studies were very diverse. Unlike BERT, the limited access to GPT-2 appears to have restricted the languages it was modelled in as well as the applicable mitigation techniques.

### **GPT-3**

In 2020, OpenAI released GPT-2's subsequent version GPT-3 [12]. GPT-3 is more than 10 times bigger than GPT-2 or auto-regressive models existed before it with an architecture hosting over 175 billion parameters. Unlike in GPT-2, the transformer nodes in GPT-3

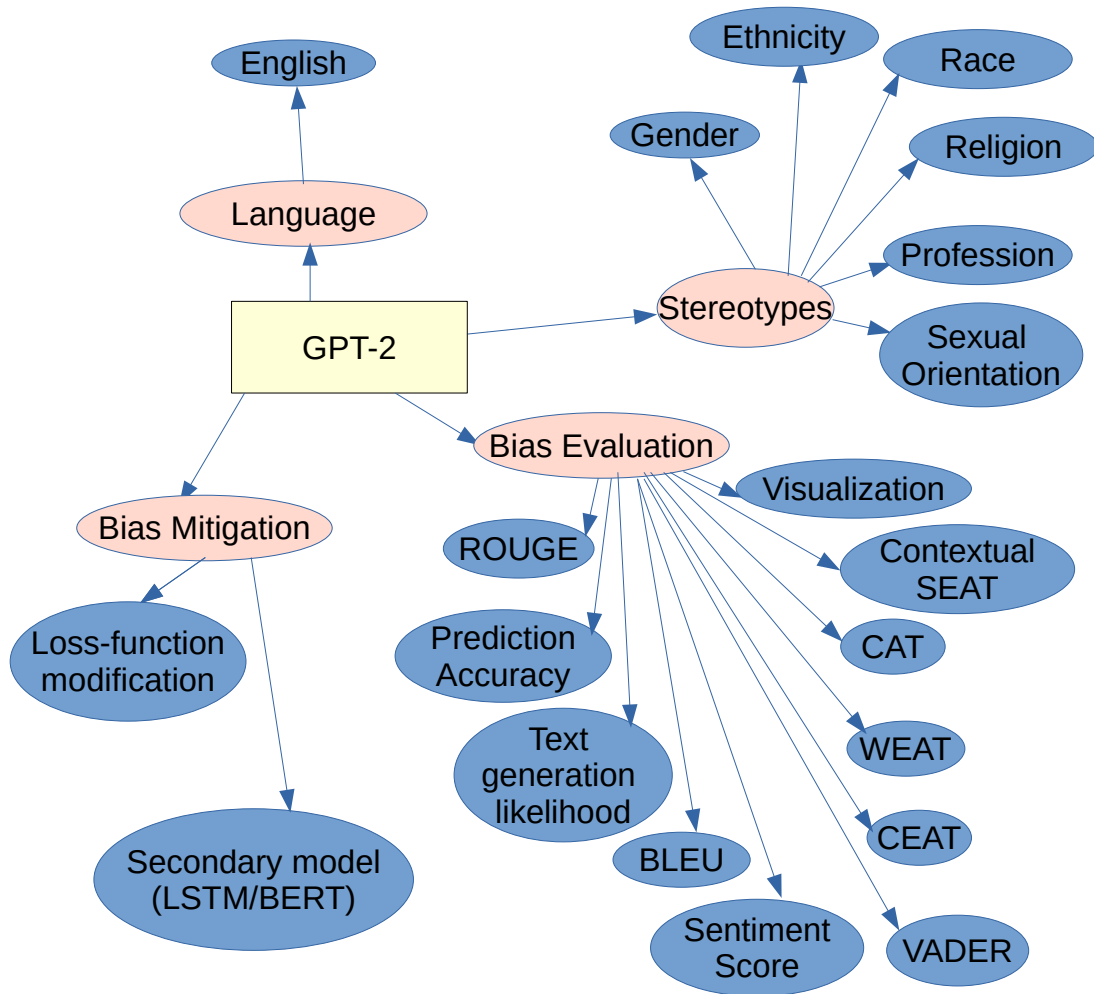


Figure 2.6: Summary of bias analyses done in GPT-2 through based on published papers.

use alternating dense and locally banded sparse attention patterns. Table 2.8 enlists papers that analyze bias in GPT-3.

## GPT-4

Unlike its previous versions, GPT-4 [13] is a multimodal model which takes text as well as image inputs and outputs texts. Released on March 2023, GPT-4 has shown human level excellence on various professional and academic benchmarks. For example, GPT-4 can pass the Bar examination with a score on the 90<sup>th</sup> percentile, while its predecessor GPT-3.5 (variant of GPT-3) was only capable of scoring in the 10<sup>th</sup> percentile. OpenAI creators have put a disclaimer that GPT-4 is still less capable than humans in many real-world scenarios. However, due to its accessibility and avid use via ChatGPT [110], the

LLM	Reference	Data	Stereotype	Language	Evaluation	Mitigation
GPT-3	Brown et al. [12]	Common Craw, Web- Text2, Books1, Books2, Wikipedia	Gender, Race, Reli- gion	English	Text genera- tion	NA
	McGuffie et al. [114]	Common Craw, Web- Text2, Books1, Books2, Wikipedia	Ideological, Political, Race	English	QA, Text Generation	NA
	Floridi et al. [115]	Common Craw, Web- Text2, Books1, Books2, Wikipedia	Race	English	Text Genera- tion	NA
	Abid et al. [116]	Common Craw, Web- Text2, Books1, Books2, Wikipedia, Humans of New York images	Ethnicity	English	Analogies, associations, Text Genera- tion	Positive Contextual- ization

Table 2.8: Papers analyzing bias in GPT-3.

biases present in it might cause adverse effects in the society.

## ChatGPT

ChatGPT [110] is a conversation user interface with GPT-4 in its backend. This tool is available for free at <https://chat.openai.com/>. Previously, Google [117] was the technology widely used to answer queries or give instructions. While Google only sorts the information it finds in the web by ranking the websites, ChatGPT provides the answer to the prompt it was given and sometimes also explains the answer it provides. Figure 2.7 shows an example of prompt given to ChatGPT and the output it provided. The lack of referencing in its responses makes it difficult to know where it is drawing its answers from. Its responses could be a culmination of various kinds of biases that will remain unknown and difficult to trace as well as quantify. Moreover, ChatGPT does not guarantee to provide correct answers. It might be harmful to the society when ChatGPT is used as a fact learning source, and miscommunication is spread widely. When using ChatGPT, one should be aware of its limitations and potential harm it might cause upon misutilization .



Make sentences with the word policy in different meanings



1. The policy of our company is to prioritize customer satisfaction above all else.
2. The government's policy on immigration has been a topic of much debate in recent years.
3. I just received my car insurance policy in the mail and need to review the coverage details.
4. The country's foreign policy aims to promote peace and stability in the region.
5. The social policy of providing free education to all children has helped to reduce inequality.



Figure 2.7: Example of an inquiry to ChatGPT (<https://chat.openai.com/>) and its response.

### 2.3.7 Summary of Stereotype Biases in LLMs

The use of deep learning models in NLP (LLMs) first started in 2013 with generation of word embeddings namely, Word2vec [43, 44]. However, the bias analyses on word embedding models did not start until 2016. Fig. 2.8 presents a graph showing the model complexity against year timeline for various models' bias evaluation. Here, the model complexity was calculated as the number of learnable parameters of the deep learning model.

From 2016 to 2018, only Word2vec, GloVe, and their variants were under scrutiny for unfavorable stereotype biases. In 2018, the most advanced language models like ELMo, Transformer, BERT and others were developed. Thus, in 2019, most of these models were evaluated for biases. We also see steep rise in the complexity of the models. In 2020, less number of LLMs were examined for bias. 2020 was also the year when OpenAI's largest model GPT-3 was released. Similar to previous GPT versions, the fully trained GPT-3 model is not publicly available but a smaller version of it can be used via an API (Application Programming Interface). Therefore, testing various inputs was only possible for analyzing GPT-3's bias starting from 2020. We also observe that 2020 onward the original transformer model is no longer examined for bias. A reason for this is the rising popularity of its descendants BERT and GPT variations.

The six most frequent LLM bias literature were Word2vec, GloVe, CBOW, BERT, GPT-2 and GPT-3. There were 11 different unfavorable stereotype biases mentioned in

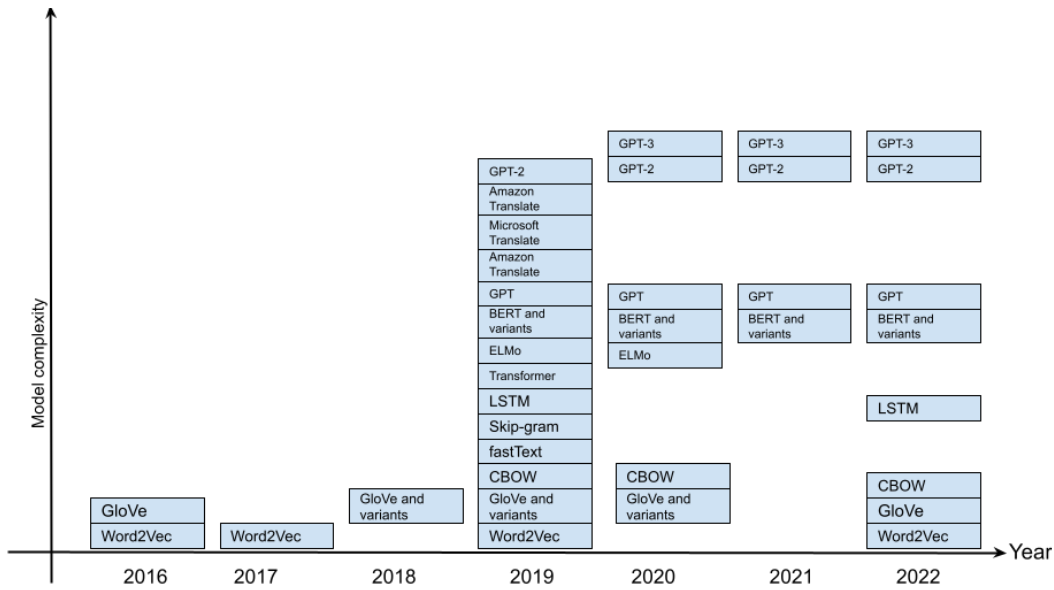


Figure 2.8: Graph of complexity of models over time for bias analyses based on published papers.

the literature namely general, gender, ethnicity, religion, age, race, profession, sexual orientation, disability, political and ideological. Fig. 2.9 shows the association of different models with different types of stereotype biases. The width of the arrows signify the number of papers that displayed a particular type of bias for a model. Thus, the wider the width of the arrow, the more number of papers are published proving that association.

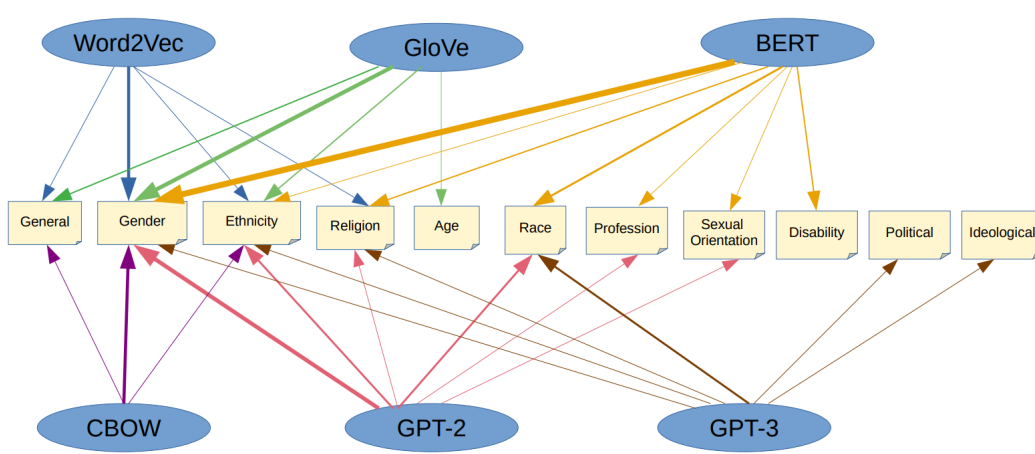


Figure 2.9: Graph of various bias analyses for the popular LLMs based on published papers.

As observed in the previous tables, Fig. 2.9 also shows gender as the most common stereotype bias reviewed in popular LLMs. Out of the six models, five models had their widest arrows to gender. Thus, maximum number of publications topics dealt with gen-



der bias in Word2vec, GloVe, BERT, CBOW and GPT-2. Ethnicity was another popular model which had links to all six models. GPT-3 had the maximum number of publications found for race. The least explored type of stereotype biases with only one model associated to each are age, disability, political and ideological.

## 2.4 Bias Measures

Numerous bias measures have been proposed in the last decade. The major challenge for bias measures whether these measures actually indicate the presence or absence of bias in the LLMs. We explain one of the most common bias measure, WEAT [35], and its derivatives, and then describe another measure named as the SAME score [36] that addresses the limitations of WEAT.

### 2.4.1 WEAT and its Derivatives

The bias measure WEAT [35] was largely inspired by the Implicit Association Test (IAT) [118] which is used to measure societal stereotype biases in humans in the field of psychology. WEAT also adopted the concept of measuring bias between two target ( $X, Y$ ) and two attribute ( $A, B$ ) sets using a cosine based calculation. The target sets are 2 groups which are hypothesised to have imbalanced treatment by the NLP embedding or model. The two attribute sets are groups of words describing some characteristic, attitude, traits of the two respective target sets. High values of WEAT indicate bias is present, i.e.,  $X$  is highly associated to  $A$  whereas  $Y$  is highly associated to  $B$ . Low values of WEAT indicates the hypothesised bias ( $X$  to  $A$ ,  $Y$  to  $B$ ) is not significant. WEAT is calculated with the following two equations:

$$s(w, A, B) = \frac{1}{n} \sum_{a \in A} \cos(w, a) - \frac{1}{n} \sum_{b \in B} \cos(w, b) \quad (2.1)$$

$$\mathbf{WEAT}(X, Y, A, B) = \frac{\mathit{mean}_{x \in X} s(x, A, B) - \mathit{mean}_{y \in Y} s(y, A, B)}{\mathit{stddev}_{w \in X \cup Y} s(w, A, B)} \quad (2.2)$$

Although WEAT [35] is the most popular bias measure used in NLP literature, there are a few bias evaluation techniques which are very similar to WEAT. Table 2.9 summarizes a few methods that are very similar to WEAT namely, WEFAT (Word Embedding Factual Association Test) [35], MAC (Mean Average Cosine Similarity), XWEAT (Multilingual and cross-lingual extension of WEAT) [88], WEAT\* [86], SEAT (Sentence Encoder Association Test) [119] and CEAT (Contextualized Embedding Association Test) [89].

Bias measure	Description
WEFAT (Word Embedding Factual Association Test) [35]	<p>Introduced in the same paper as WEAT [35], WEFAT was a measure proposed to give bias score for a word <math>w</math> and two attribute sets <math>A, B</math>. The WEFAT bias was denoted by <math>s(w, A, B)</math> and computed in the following way</p> $s(w, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std} - \text{dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})} \quad (2.3)$
MAC (Mean Average Cosine Similarity) [69]	<p>The motivation behind MAC is to calculate a weighted cosine similarity. Suppose a function <math>S</math> computes the mean cosine distance between target <math>\mathbf{t}_i \in T</math> and attribute <math>A_j</math> as follows:</p> $S(\mathbf{t}_i, A_j) = \frac{1}{N} \sum_{a \in A_j} \cos(\mathbf{t}_i, \mathbf{a}) \quad (2.4)$ <p>Then, MAC (Mean Average Cosine Similarity) is defines as</p> $MAC(T, A) = \frac{1}{ T  A } \sum_{\mathbf{t}_i \in T} \sum_{A_j \in A} S(\mathbf{t}_i, A_j) \quad (2.5)$
XWEAT (Multilingual and cross-lingual extension of WEAT) [88]	<p>Instead of monolingual word embeddings space, perform WEAT[35] in Cross-Lingual embeddings (CLEs) space. They then quantified gender bias across various languages.</p>
WEAT* [86]	<p>Perform WEAT on meaningful male-vs-female associations instead of stereotypical ones. For example <math>X = \{\text{man, male, boy, brother, him, his, son}\}</math>, <math>Y = \{\text{woman, female, girl, sister, her, hers, daughter}\}</math>, <math>A = \{\text{gentleman, king, patriarchy}\}</math>, <math>B = \{\text{madam, queen, matriarch}\}</math></p>
SEAT (Sentence Encoder Association Test) [119]	<p>Use target sets <math>X, Y</math> and attribute sets <math>A, B</math> as described in WEAT but create Neutral sentence embeddings instead of word embeddings. For example if <math>X = \{\text{man}\}</math> in WEAT, it comes <math>X = \{\text{This is a man}\}</math> in SEAT.</p>
CEAT (Contextualized Embedding Association Test) [89]	<p>CEAT summarizes the magnitude of overall bias in neural language models by incorporating a random-effects model. This is done by using techniques to calculate social and intersectional biases in contextualized word embeddings.</p>

Table 2.9: Bias measures similar to WEAT

## 2.4.2 SAME

More recently, a bias metric called SAME was introduced in Schroder et al. [36]. They begin their paper by defining 4 concepts necessary for bias metrics that considers geometric (cosine) similarity of word embeddings namely: trustworthiness, comparability,

skewness and stereotype. The concepts of trustworthiness and comparability are well described but the concepts of skewness and stereotype are not adequately characterized. However, the paper claims that SAME is a good metric that considers all four concepts. With experiments on numerous models, they show that SAME metric performs the best. Equation 2.6 shows how the same metric is calculated given a target word set  $W$  and two attribute sets  $A_i$  and  $A_j$ .

$$SAME(W, A_i, A_j) = \frac{1}{|W|} \sum_{w \in W} |\cos(\mathbf{w}, \hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)| \quad (2.6)$$

### 2.4.3 Limitations of WEAT and SAME

WEAT has been used or analyzed on multiple language models and text corpora [77, 85, 88, 87, 86, 89, 120, 121, 18, 95, 122]. May et al. [119] warn us that using WEAT as the ultimate measure of bias might not be ideal. They instead propose using SEAT (Sentence Embedding Association Test), which is WEAT’s adaptation for contextualised sentence embedding. Theoretical flaws of WEAT are also highlighted in Ethayarajh et al. [42] via comprehensive mathematical proofs. Although the theoretical and geometrical concepts of trustworthiness and comparability are well defined in the SAME [36] metric, our experiments show that all types of biases are not always best captured by the metric.

Unlike most of the research works that focus only on measuring unfavorable societal stereotype bias, this dissertation will also focus on the performance of bias evaluation measures on categorical bias that is necessary for the model to learn. We first conducted experiments on last four hidden layers of BERT (*bert-base-uncased*) to find out whether WEAT and SAME is capable of detecting categorical bias. We use facts such as country to its capitals with provided with correct and incorrect pairs (Table 2.10). We expect the model to learn country capital facts from its training text corpus such as Wikipedia and BookCorpus. Therefore, presence of learning these facts should show high values for categorical bias. WEAT values were always very high (either 1.999 or 2.000) whether we provided correct or incorrect country capitals. The SAME score showed low values (range

[0.041 , 0.063]) for correct capitals and highest values (0.184 and 0.164) for incorrect capitals. Both WEAT and SAME scores are incapable of identifying categorical biases present in the model. Therefore, using these bias evaluation measures for quantifying stereotype bias may not be appropriate.

Bias type	X	Y	A	B	WEAT	SAME
Correct Capital	'Italy'	'Rome'	'Germany'	'Berlin'	2.000	0.062
	'Italy'	'Rome'	'France'	'Paris'	2.000	0.063
	'Germany'	'Berlin'	'France'	'Paris'	1.999	0.053
Incorrect Capital	'Italy'	'Berlin'	'Germany'	'Rome'	2.000	0.031
	'France'	'Bern'	'Switzerland'	'Paris'	2.000	0.168
	'Germany'	'Paris'	'France'	'Berlin'	1.999	0.046

Table 2.10: Quantifying categorical biases using WEAT and SAME

Although the results in Table 2.10 showed that WEAT and SAME is not capable of detecting categorical bias, we also conducted our experiments on gender bias examples. Table 2.11 shows the values WEAT and SAME score have for the gender bias experiments. WEAT values are either too high (1.999 or 2.000) like before or too low (-1.999). These WEAT values are very extreme and show no variation again. SAME scores display some variation but all of them are close to 0. Due to its inconsistency for categorical bias, we cannot rely on the values we get for stereotype bias using WEAT and SAME. We require a bias measure that can appropriately quantify categorical and therefore, stereotype bias appropriately.

Bias type	X	Y	A	B	WEAT	SAME
Gender	'man'	'woman'	'doctor'	'nurse'	2.000	0.022
	'man'	'woman'	'child'	'beautiful'	-1.999	-0.022
	'man'	'woman'	'engineer'	'homemaker'	1.999	0.121
	'he'	'she'	'doctor'	'nurse'	2.000	0.028

Table 2.11: Quantifying gender bias using WEAT and SAME

## 2.5 Quantify Model Tendency after Customization

In order to make models suitable for their domain applications, it is common to apply transfer learning techniques for customization. Transfer learning is a popular method

where a pre-trained model is used as starting foundational model for a different but related task. LLMs like BERT [10] and GPT-2 [102] were trained on enormous text corpora and high computing resources that are not available to all. By using these publicly available model for customization on domain, one saves computing resources, financial resources and time.

Majority of the research work that customizes LLMs on domain text corpus are tested for customization progress on a particular task. To monitor improvement on these tasks, already existing measures like accuracy, precision, recall, correlation coefficient and F1-scores are used to quantify model tendency upon customization. Navgi et al. [123], Peng et al. [124], Qasim et al. [125], and Kang et al. [126] do not devise a new way to quantify domain knowledge and use previously mentioned performance metrics that are used to monitor classification task performance.

Another way the shift in model tendencies are observed are in embedding neighbors. Bogust et al. [127] introduced an interactive tool called Embedding Comparator that allows for the global comparison of two model embeddings by visualizing their local neighborhoods for a given word or emoji using k-nearest neighbors. This tool is particularly effective for comparing fixed embeddings such as LSTM, Word2vec, or GloVe. However, it is not suitable for determining whether contextual learning models such as BERT are acquiring domain knowledge. One approach for assessing the acquisition of domain knowledge by a model for words with multiple meanings is to partition the dimensions of BERT embeddings based on their various meanings using techniques such as K-means [128] or k-NN [129]. To the best of our knowledge, there is still lack of research on how to quantify the extent to which a model has acquired domain knowledge that is independent of the NLP task.

## 2.6 Summary

In this chapter, after providing an overview of bias studies in the literature, the unfavorable stereotype bias research focusing various LLMs were explained. Moreover, common bias measures such as WEAT and the SAME score were explained with their limitations.

Finally, we have highlighted the lack of generalizable measures to quantify the domain bias. In a similar vein, class level bias research has not been addressed as the unfavorable stereotype bias in the literature.

# Chapter 3

## Directional Pairwise Class Confusion Bias for Evaluating Class Level Bias

Recent advances in Natural Language Processing have led to powerful and sophisticated models like BERT (Bidirectional Encoder Representations from Transformers) [10] that have bias. These models are mostly trained on text corpora that deviate in important ways from the text encountered by a chatbot in a problem-specific context. While a lot of research in the past has focused on measuring and mitigating bias with respect to protected attributes (stereotyping like gender, race, ethnicity, etc.), there is lack of research in model bias with respect to classification labels. In this chapter, we investigate whether a NLP classification model hugely favors one class with respect to another. We propose a **class level bias evaluation method** called *directional pairwise class confusion bias* that highlights the chatbot intent classification model's bias on pairs of classes. We also present a few strategies on how our directional pairwise confusion bias could be utilized to develop mitigation methods for biased pairs.

This chapter includes a substantial amount of text, figures and tables from papers whose primary author is the same as this dissertation, and published by IEEE and World Scientific in the following research papers:

- © 2022 IEEE. Reprinted, with permission, from Sayenju, Sudhashree, Ramazan Aygun, Jonathan Boardman, Duleep Prasanna Rathgamage Don, Yifan Zhang, Bill

Franks, Sereres Johnston, George Lee, Dan Sullivan, and Girish Modgil. "Directional pairwise class confusion bias and its mitigation." In 2022 IEEE 16th International Conference on Semantic Computing (ICSC).

- *Quantification and Mitigation of Directional Pairwise Class Confusion Bias in a Chatbot Intent Classification Model* by Sayenju, Sudhashree, Ramazan Aygun, Jonathan Boardman, Duleep Prasanna Rathgamage Don, Yifan Zhang, Bill Franks, Sereres Johnston, George Lee, Dan Sullivan, and Girish Modgil. In International Journal of Semantic Computing: 497-520. © 2022 World Scientific.

### 3.1 Motivation

Conversational chatbots are commonly used by businesses to help end users or customers with their concerns or problems to provide immediate assistance during anytime of the week. With the help of new methods in Artificial Intelligence (AI) and Natural Language Processing (NLP), chatbots aim to provide better customer service. Using chatbots, companies save time and financial resources by utilizing their human resources for more complicated tasks. Additionally, chatbots are also convenient for customers since they do not have to read through large FAQ pages or be in the waiting list until a customer support employee is available.

Chatbots first aim to find the intent behind human text utterances. After recognizing the intent, chatbots can provide appropriate information or guide the end users in the correct direction. Although chatbots have evolved over time they still have some limitations [130, 131]. Chatbots are mostly trained on a specific domain. Therefore, if a customer asks regarding a slightly different topic, it might not know how to respond like a human. Chatbots are generally incapable of recognizing grammatical errors or misspellings. In cases where customers come from different backgrounds, chatbots might not be able to understand their accents or lingo. This leads to poor conversation understanding and runs the risk of incorrect intent classification. Such biases in intent classification could cause chatbots to give replies that sound robotic, give ambiguous answers, lead customers in the



wrong direction or even frustrate the customer [132].

In general machine learning models are vulnerable to bias. As a result of this, their decision could be undesirable or unfair. To understand how bias occurs in machine learning models, it is important to recognize the intimate relationship between bias in data and bias in algorithms. Since most machine learning models are data-driven, it is possible for these models to learn the bias in data during training and reflect it in predictions. Also, algorithms can change the level of bias in data or display a bias that is not present in data. The outcome of such biased models is then introduced to real-world systems such as chatbots and subsequently affects human decisions. Then it may produce even more biased data for future model training [19].

The chapter focuses on class level bias of a NLP classification model. The results presented in this chapter shows the bias of a BERT [10] model used in a chatbot for intent classification. We propose a measure called *directional pairwise class confusion bias*. The aim of this measure is to find whether the trained model makes mispredictions in the favor of one class compared to another class. The directional pairwise class confusion bias is visualized to reveal the most critical bias cases. Such biases in the model might arise due to class imbalance in the training data, or other semantic biases encapsulated through accents, misspelling or chatbot's limited domain knowledge. Additionally, this chapter also proposes a few strategies on how the directional pairwise confusion bias could be utilized to develop mitigation methods for biased pairs.

## **3.2 Building Intent Classification Model using Transfer Learning with BERT Model**

Intent classification with more than two classes is a complex task requiring a highly developed model. In order to have high predictive power in our model and save time with the limited computing resources, transfer learning was applied on a pre-trained BERT (Bidirectional Encoder Representations from Transformers [10]) model namely, *bert-base-uncased*. The fundamental concept of transfer learning is to reuse a machine learning

	Training Loss	Valid. Loss	Valid. Accur.	Training Time	Validation Time
<b>epoch</b>					
<b>1</b>	0.67	0.51	0.85	0:42:37	0:01:39
<b>2</b>	0.43	0.48	0.85	0:42:38	0:01:39
<b>3</b>	0.35	0.50	0.85	0:42:37	0:01:39
<b>4</b>	0.28	0.52	0.86	0:43:26	0:01:40

Figure 3.1: Results of all 4 epochs in the training phase. © 2022 IEEE

model originally developed for one task in a different task with limited dataset.

The pre-trained model *bert-base-uncased* was trained on BooksCorpus and English Wikipedia (excluding lists, tables and headers). As the name suggests *bert-base-uncased* was trained on lower-cased English text. The model consists of 12 transformer blocks, 768-hidden layers, 12 self attention-heads and a total of 110 million parameters. Training of *bert-base-uncased* required total of 16 TPUs. Transfer learning was done on *bert-base-uncased* for the variant that does single sentence classification task. The input for training was our chatbot data (human utterances) and their corresponding class labels were used in the softmax layer. The 4 epochs were run giving validation accuracy of 86% in epoch 4 (Fig.3.1).

Once training was complete, the model was evaluated on an unseen test set. The test accuracy of the model was 74.4%. Considering there are 18 classes, the test accuracy is fairly good and performs much better than a random guess ( $\frac{1}{\text{number of classes}} = \frac{1}{18} = 5.5\%$ ). However, note that the goal of our research in this chapter is not to improve the performance of the model but to investigate model's bias at class level. This model will be used to analyze bias.

### 3.3 Directional Pairwise Class Confusion Bias

The most common measure to evaluate a machine learning model is *accuracy*. However, *accuracy* cannot provide insight about the model's performance if the class distribution is unbalanced. Hence, measures such as *precision*, *recall*, *sensitivity*, and *specificity* could

be used for evaluating the model at the class level. Still these measures do not provide where the mispredictions originate from at the class level. For example, the sensitivity measure may not reveal the misclassifications that happen with respect to a specific class. Hence, the model could be biased towards one class when the actual instances belong to another class.

We begin our bias analysis by plotting the confusion matrix generated by the fine-tuned BERT model on the test dataset. Fig.3.2 shows the confusion matrix plot with true labels on the rows and model predicted labels in the columns. The values in each cell represent the number of samples that were predicted as the column label for the correct row label. Since there are a lot of classes, looking at the confusion matrix with a naked eye might not highlight the prominent values. Fig. 3.3 shows a heatmap of the confusion matrix. The largest (dark blue) values are found in the diagonal. This confirms the model being mostly accurate (74.4%).

In Fig. 3.2 there are cells above and below the diagonal which have values greater than 0. Those cells show bias at class level and are of interest for this research. Since classes were not distributed evenly, it is difficult to observe any biases directly from the confusion matrix. If there is any bias, quantifying the bias is essential to prioritize bias mitigation.

In order to visualize the biases more clearly, the confusion matrix was modified to highlight the bias between a pair of classes. Since typically the classes are unbalanced, the confusion matrix needs to be normalized. Each cell in the confusion matrix (Fig. 3.2) was divided by the maximum of its column.

$$C'(i, j) = \frac{C(i, j)}{\max_{k=1, \dots, n} C(k, j)} \quad (3.1)$$

where  $C$  represents the confusion matrix,  $C(i, j)$  represents the number of classifications predicted class  $c_j$  but whose ground truth was  $c_i$ , and  $C'$  indicates the normalized confusion matrix. Doing this operation converts all the values in the matrix between 0 and 1. Normalization could be done with respect to rows (actual labels) rather than columns (predictions). Since the user of a machine learning model observes the predictions, it makes more sense to normalize with respect to the predictions. The normalized results

	Account_Related	Billing_Related	Cancel_Related	Claim_Related	Coverage_Related	Discount_Related	Document_Related	Escalation	EverythingElse	Payment_Related	Policy_Related	Premium_Related	Quote_Related	SmallTalk	deny
Account_Related	182	12	1	0	0	1	15	9	12	8	23	0	0	3	0
Billing_Related	1	320	14	2	0	8	10	3	2	68	5	14	1	1	0
Cancel_Related	2	14	1627	5	0	1	11	5	1	14	36	0	3	4	0
Claim_Related	0	3	1	376	9	1	22	6	6	4	3	3	8	0	0
Coverage_Related	0	2	6	25	234	2	127	3	4	3	30	4	46	0	0
Discount_Related	2	2	1	0	0	498	10	13	2	1	4	4	3	1	0
Document_Related	13	5	15	38	42	13	1964	25	32	32	111	23	42	7	0
Escalation	6	9	14	5	0	25	31	1611	50	30	27	8	6	40	0
EverythingElse	9	4	6	11	4	8	35	65	422	11	50	2	11	44	0
Payment_Related	5	51	12	2	3	7	37	32	14	2617	23	55	8	4	0
Policy_Related	21	2	45	3	30	3	116	20	32	28	2625	3	71	7	0
Premium_Related	1	15	7	3	9	7	16	2	1	57	10	423	23	0	0
Quote_Related	2	1	3	1	34	6	75	3	7	7	75	13	1083	5	0
SmallTalk	4	1	2	4	3	3	11	44	58	3	13	2	5	924	0
deny	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

Figure 3.2: Class Confusion matrix © 2022 IEEE

are visualized in Fig. 3.4. As the original model is highly accurate (74.4%), the diagonal elements have the highest values in their respective columns. Due to this, most of the diagonal elements have the largest value 1.

The diagonal elements in Fig. 3.4 are accurate predictions and do not show bias. Hence,  $C'$  matrix is updated as  $C'(i, i) = 0$  for every value in its diagonal. The cells that have some degree of blue color above and below the diagonal show cases where bias is present. In order to visualize these cases more clearly, the values in the diagonal were muted by setting them to be 0. The result showing the bias pairs are then visualized in Fig. 3.5. Equation 3.1 is updated as follows:

$$identity(i, j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (3.2)$$

$$C'(i, j) = \frac{C(i, j) * (1 - identity(i, j))}{\max_{k=1, \dots, n} C(k, j)} \quad (3.3)$$

We coined the term *directional pairwise class confusion bias* for evaluating the bias. This indicates the likelihood of classifying an instance in one specific class into another

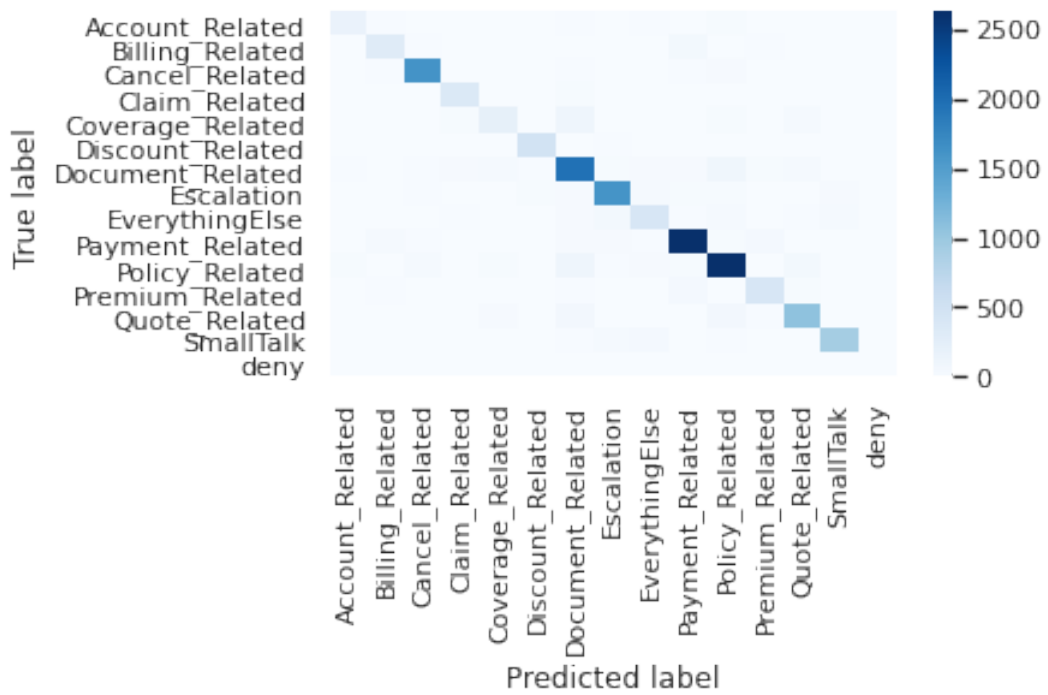


Figure 3.3: Heatmap of Class Confusion matrix © 2022 IEEE

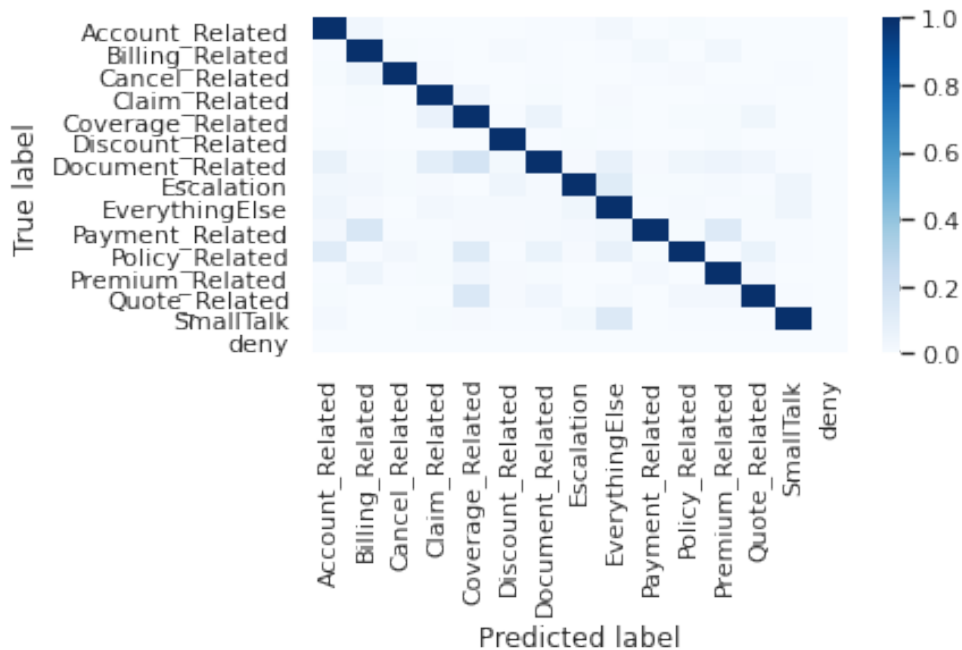


Figure 3.4: Dividing each cell in the confusion matrix by the maximum of its column. © 2022 IEEE

class. Thus, there is a direction of misclassification.

Although Fig. 3.5 gives a clear view of directional pairwise class confusion bias, we are only interested in cases where the bias is strongly present. To reflect on the cases where bias is strong, the directional pairwise class confusion bias matrix was further pruned by setting a threshold. The threshold filter will return only those rows and columns where one of their values is above the threshold. Fig. 3.6 is the pruned matrix reflecting bias cases above threshold of 0.15. The plot clearly shows a strong bias for cases which are Coverage\_Related but were classified by our BERT model as Document\_Related. Now, we may formally *define directional pairwise class confusion bias*.

### 3.3.1 Definition: Directional Pairwise Class Confusion Bias

$c_i \xrightarrow{b} c_j$  represents a directional pairwise bias from class  $c_i$  to  $c_j$  for a machine learning model that indicates that there is a likelihood of a sample belonging to class  $c_i$  being classified as  $c_j$  by the trained model. This bias is quantified as  $\beta(c_i \xrightarrow{b} c_j) = C'(i, j)$  and this bias is considered to be significant if  $\beta(c_i \xrightarrow{b} c_j) > \theta_b$  where  $\theta_b$  is a threshold for significance of bias and determined by an expert. The antecedent is called as the *source bias class* whereas the consequent is called as the *destination bias class*.

Directional pairwise class does not have the *identity property*. In other words,  $\beta(c_i \xrightarrow{b} c_i) = 0$ . The *symmetry* property may not always hold. Thus, if  $c_i \xrightarrow{b} c_j$  is true, we cannot infer that  $c_j \xrightarrow{b} c_i$ . Similarly, we cannot claim the *anti-symmetry* property, if  $c_i \xrightarrow{b} c_j$  is true, there is a likelihood of  $c_j \xrightarrow{b} c_i$ . The *transitive* property is unlikely to hold, since  $c_i \xrightarrow{b} c_j$  and  $c_j \xrightarrow{b} c_k$ , there is no guarantee that  $c_i \xrightarrow{b} c_k$  is true.

Fig. 3.6 shows the strongest pairwise class confusion bias between pair (*Coverage\_Related*, *Document\_Related*) represented as  $c_{coverage} \xrightarrow{b} c_{document}$ . Note that other pairs like  $c_{billing} \xrightarrow{b} c_{payment}$ ,  $c_{coverage} \xrightarrow{b} c_{quote}$ ,  $c_{everythingElse} \xrightarrow{b} c_{escalation}$ ,  $c_{everythingElse} \xrightarrow{b} c_{document}$  and  $c_{quote} \xrightarrow{b} c_{document}$  also show significant bias.

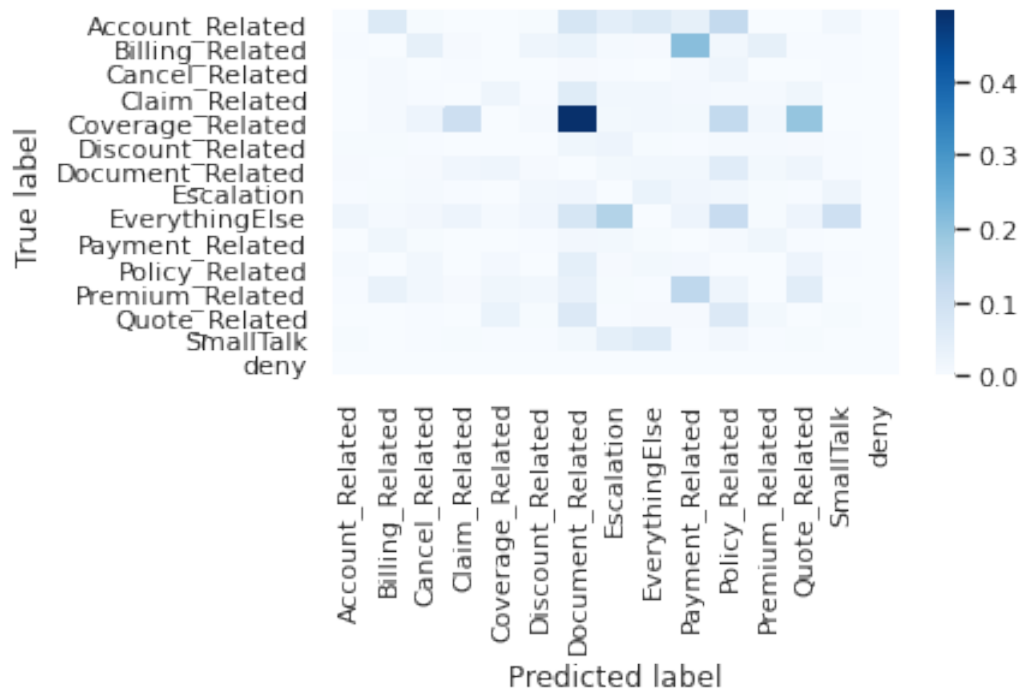


Figure 3.5: Directional Pairwise Class Confusion Bias © 2022 IEEE

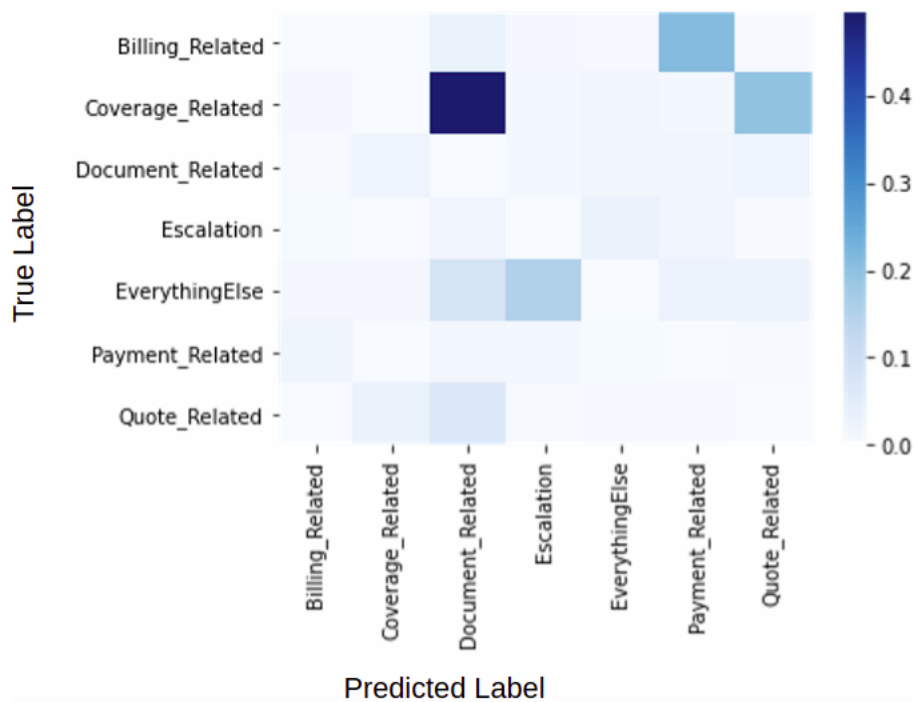


Figure 3.6: Pruned Directional Pairwise Class Confusion Bias matrix after threshold was set to 0.15. © 2022 IEEE

### 3.3.2 Bias Mitigation Process

The main purpose of defining and quantifying our bias was to find ways of mitigating it. Bias mitigation plays a crucial role in ensuring fairness of models. Before mitigation, it is essential to know the original model's behavior. Thus, the first step is to analyze the classification results from the original model. In presence of directional pairwise class confusion bias, the model shows a high tendency to mispredict instances of class  $c_s$  as  $c_d$  instead. We assume the original model to be a multi-class model which is capable of distinguishing a large number of classes. However, creating a globally accurate model to separate all the classes might overlook pairwise mispredictions. We will demonstrate our mitigation processes using sample source ( $c_s$ ) and destination ( $c_d$ ) classes where  $c_s \xrightarrow{b} c_d$  exists. Directional pairwise bias occurs if the model predicts a large fraction of the test instances to be of class  $c_d$  when the ground truth was  $c_s$ .

We present two sample strategies for mitigation to show how directional pairwise confusion bias could be utilized: *priori bias mitigator* and *posteriori bias mitigator*. Both the mitigation techniques *priori bias mitigation* and *posteriori bias mitigator* are very similar. The idea of these bias mitigation techniques is to use a secondary binary classifier whenever directional pairwise class confusion bias is observed, so that the secondary model learns to distinguish the source bias class from the destination bias class. The secondary classifier could increase the recall of  $c_s$  using a binary classification model designed to better distinguish  $c_d$  and  $c_s$ . Note that the original model favors  $c_d$  compared to  $c_s$ , hence the recall of  $c_s$  turns out to be lower.

The main difference between the priori and posteriori bias mitigator is the selection of the training set to build the secondary classifier. In our experiments, the secondary model is a binary Random Forest classifier. During evaluation, the destination class ( $c_d$ ) predictions are separated for the secondary model. We feed  $c_d$  predicted instances from the original model into the secondary model, in hopes of getting less mispredictions. The secondary model's predictions (now either  $c_d$  or  $c_s$ ) are merged with the remaining test predictions for mitigation evaluation. Lastly, the original model's results are compared with the priori bias mitigation results (Fig. 3.7).



*Priori Bias Mitigator.* The training set for the *priori bias mitigator* is the union of samples that belong to classes  $c_d$  and  $c_s$  from the original training set. The *priori bias mitigator* has access to the all the samples of  $c_d$  and  $c_s$  from the training set.

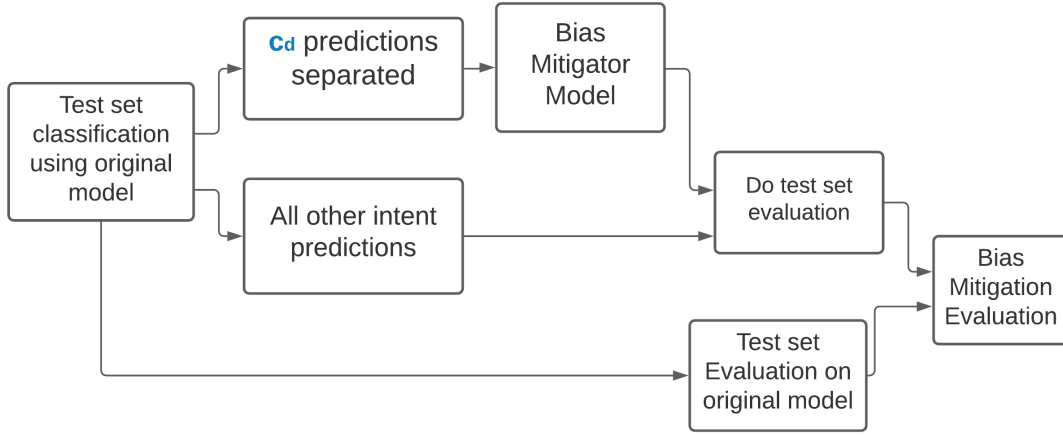


Figure 3.7: Evaluation of bias mitigation process © 2022 IEEE

*Posteriori Bias Mitigator.* Both *priori bias mitigator* and *posteriori bias mitigator* are both trained based on the subsets of the original training set. While *priori bias mitigator* takes the projection of selected classes, *posteriori bias mitigator* selects the training set based on the predictions of the original classifier for the training set. The idea of this method is to train the secondary model based on  $c_d$  and  $c_s$  predictions of the original model. This is why we call it as the secondary classifier, *posteriori bias mitigator*.

## 3.4 Experiments

### 3.4.1 Data

The dataset and label sets used in our experiments are provided by Travelers Indemnity Company. This dataset consists of customer (human) utterances with intent class. In total there were 128,201 user utterances, each belonging to one of 21 classes. A subset of the data was separated for modelling into training and testing. The training set had 96,150 utterances and 18 classes. The test set had 20,031 utterances with 15 labels.

We investigate whether either of the the mitigators helps us mitigate the directional

pairwise class confusion bias observed for our chatbot intent classification model in Section 3.3. We demonstrate the *priori bias mitigator* and *posteriori bias mitigator* experiments on three biased pairs:  $c_{coverage} \xrightarrow{b} c_{document}$ ,  $c_{billing} \xrightarrow{b} c_{payment}$  and  $c_{everythingElse} \xrightarrow{b} c_{escalation}$ . We use recall to analyze before and after effects of mitigation. We choose recall since it demonstrates the ability of the model to retrieve relevant cases within a dataset.

### 3.4.2 Bias in the Original BERT Model for Chatbot’s Intent Classification

	precision	recall	f1-score	support
Account_Related	0.73	0.68	0.71	266
Billing_Related	0.72	0.71	0.71	449
Cancel_Related	0.93	0.94	0.94	1723
Claim_Related	0.78	0.85	0.81	442
Coverage_Related	0.63	0.49	0.55	475
Discount_Related	0.85	0.92	0.89	541
Document_Related	0.79	0.83	0.81	2363
Escalation	0.87	0.87	0.87	1862
EverythingElse	0.65	0.62	0.64	682
Payment_Related	0.90	0.91	0.90	2870
Policy_Related	0.85	0.87	0.86	3017
Premium_Related	0.75	0.74	0.74	574
Quote_Related	0.82	0.82	0.82	1315
SmallTalk	0.88	0.86	0.87	1077
deny	0.00	0.00	0.00	1

Figure 3.8: Test set evaluation on original intent classification BERT model without mitigating bias.

Before mitigation, it is important to analyze the original state of the chatbot intent classification model in terms of performance for each class. The classification report of the original BERT model are shown in Fig. 3.8. The directional pairwise bias  $c_{coverage} \xrightarrow{b} c_{document}$ ,  $c_{billing} \xrightarrow{b} c_{payment}$  and  $c_{everythingElse} \xrightarrow{b} c_{escalation}$  are framed. The recall of the pairs are highlighted in green. The difference between the recall of  $c_{coverage}$  (0.49) and  $c_{document}$  (0.83) has a large margin. The bias  $c_{billing} \xrightarrow{b} c_{payment}$  shows less extreme discrimination. The recall of  $c_{billing}$  (0.71) is less than  $c_{payment}$  (0.91). The third pair  $c_{everythingElse} \xrightarrow{b} c_{escalation}$  shows similar difference between recall of  $c_{everythingElse}$  (0.62) and  $c_{escalation}$  (0.87). If the

directional pairwise bias is mitigated by a significant level, we expect to see improvement in the recall for the source class (lower performing class in the biased pair).

### 3.4.3 Priori Bias Mitigator

The experiments conducted for *priori bias mitigator* was conducted on biased pairs:  $c_{coverage} \xrightarrow{b} c_{document}$ ,  $c_{billing} \xrightarrow{b} c_{payment}$  and  $c_{everythingElse} \xrightarrow{b} .$  The destination class  $c_d$  predictions were separated and reclassified using the bias mitigator model. The bias mitigator model used here is a binary Random Forest classifier capable of distinguishing  $c_d$  from source class  $c_s$ . The bias mitigator was trained on a subset of the original model’s training set whose ground truth were either  $c_d$  or  $c_s$ . The performance of  $c_s$  classes after using priori bias mitigator are highlighted in Table 3.1. This table summarizes the recall for the two biased classes before and after mitigation. There is increase in recall for the source biased class  $c_{coverage}$  (0.49 to 0.53) and  $c_{billing}$  (0.71 to 0.73). We saw improvement in two of the three source class recall values. This method might work for a larger dataset with a more extreme directional pairwise bias.

	$c_{coverage}$		$c_{billing}$		$c_{everythingElse}$	
	Recall	Precision	Recall	Precision	Recall	Precision
Original BERT model	0.49	0.63	0.71	0.72	0.62	0.65
$c_d$ for training secondary model	<b>0.53</b>	0.59	<b>0.73</b>	0.70	0.62	0.65

Table 3.1: Priori Bias Mitigation: Before (original BERT model) and after ( $c_d$  for training secondary model) performance of source classes  $c_{coverage}$ ,  $c_{billing}$ ,  $c_{everythingElse}$ . Improvement is seen in the recall of source classes  $c_{coverage}$  and  $c_{billing}$  (in bold font).

### 3.4.4 Posteriori Bias Mitigator

Here we mitigate the same 3 biased pairs consecutively, namely  $c_{coverage} \xrightarrow{b} c_{document}$ ,  $c_{billing} \xrightarrow{b} c_{payment}$ , and  $c_{EverythingElse} \xrightarrow{b} c_{Escalation}$ . The results of posteriori bias mitigator are shown in Table 3.2. The first biased pair mitigated was  $c_{coverage} \xrightarrow{b} c_{document}$ . The recall of our source class escalated from 0.49 to 0.54. The second pair we tested for the posteriori bias mitigator was on  $c_{billing} \xrightarrow{b} c_{payment}$ . The recall of our source class  $c_{billing}$

increased compared to the original model (0.71 to 0.73). The posteriori bias mitigator was further tested on a third biased pair  $c_{everythingElse} \xrightarrow{b} c_{Escalation}$ . The recall of our source class  $c_{everythingElse}$  increases from original model (0.62 to 0.67). We saw improvement in all of the three source class recall values.

	$c_{coverage}$		$c_{billing}$		$c_{everythingElse}$	
	Recall	Precision	Recall	Precision	Recall	Precision
Original BERT model	0.49	0.63	0.71	0.72	0.62	0.65
$c_d$ for training secondary model	<b>0.54</b>	0.57	<b>0.73</b>	0.46	<b>0.67</b>	0.41

Table 3.2: Posteriori Bias Mitigation: Before (original BERT model) and after ( $c_d$  for training secondary model) performance of source classes in biased pairs. Largest Recall for each class are written in bold font.

### 3.4.5 Discussion

Due to lack of past research on class related model bias, our research devises a way to quantify class related model bias called *directional pairwise class confusion bias*. In addition, we present two mitigation techniques to show how this bias measure could be leveraged.

The first mitigation technique we propose is called *priori bias mitigator* that uses all samples of biased classes during training. On the other hand, the posteriori bias mitigator exclusively learns from the training instances that were predicted to be either source class or destination class. This way the mitigator model adjusts itself based on the predictions of the original model. The three pairs  $c_{coverage} \xrightarrow{b} c_{document}$ ,  $c_{billing} \xrightarrow{b} c_{payment}$  and  $c_{everythingElse} \xrightarrow{b} c_{Escalation}$  were consecutively mitigated in the order presented. These bias mitigation techniques are designed to increase the recall of less favored class. However, this could lower the precision of the source class and recall of the destination class. The recall of all the experiments are summarized in Table 3.3. Among the two mitigators posteriori gave the highest recall for all source classes at a cost of precision. To mitigate bias, we suggest both methods to be evaluated to determine the best secondary model on a novel dataset. The final decision should be determined based on the tradeoff between

recall gain and precision loss.

	$c_{coverage}$		$c_{billing}$		$c_{everythingElse}$	
	Recall	Precision	Recall	Precision	Recall	Precision
Original BERT model	0.49	0.63	0.71	0.72	0.62	0.65
Priori Bias Mitigation	0.53	0.59	<b>0.73</b>	0.70	0.62	0.65
Posteriori Bias Mitigation	<b>0.54</b>	0.57	<b>0.73</b>	0.46	<b>0.67</b>	0.41

Table 3.3: Priori and Posteriori Bias Mitigation: Performance of source classes in biased pairs on all experiments. Largest Recall for each class are written in bold font.

Our experiment results could be further improved by using alternate classifiers (instead of Random Forest) for the bias mitigator model. The performance of this model might have been limited because this is a simpler model than the BERT model used for training the original model. Alternatively, creating a bias mitigator model trained on all mispredictions of the training set might work. If such mitigation does not generate robust results, another idea would be to examine the labelling of the data. There might be a hierarchy in the classification labels causing them to have directional pairwise bias. Taking our first example biased pair  $c_{coverage} \xrightarrow{b} c_{document}$ , this bias might exist because the model suggests  $c_{document}$  is a superclass of  $c_{coverage}$ . Thus, an alternate labeling technique or hierarchical classification can be implemented to distinguish the intent classes better.

### 3.5 Summary

Class imbalance and noise in text semantics can cause a NLP classification model to prefer a certain class over another. In the past, most of the research in bias of NLP models refers to human stereotyping bias or solely addresses the class imbalance problem. Therefore, we propose *directional pairwise class confusion bias*, a technique to indicate a model’s favoring of a class compared to another class. We showed how to quantify and visualize this bias to reveal heavily biased pairs. Additionally, we propose two strategies to mitigate it: *priori bias mitigator* and *posteriori bias mitigator*. The two mitigation techniques use a secondary classifier in their process to correct the biased outputs. For situations where mitigation is not as effective as it was in our data, directional class confusion bias still

provides insights about the cases that are hindering the performance of the model.

## Chapter 4

# Differential Cosine Bias Measure for Evaluating Stereotype and Categorical Bias

A vast range of Natural Language Processing (NLP) systems that are in use today have direct impact on humans. While machine learning models are expected to automatically infer world knowledge from historical texts, we should be aware not to let NLP applications consume undesired societal stereotype bias back. Many bias evaluation measures have been designed and experimented to check whether unfavorable stereotype bias is present in the model or not. Upon performing various experiments, we found out that the most popular bias measures do not always indicate bias accurately. In addition to these experimental findings, we also propose our novel *Differential Cosine Bias* measure with examples of unfavorable stereotype biases as well as necessary categorical bias that is based on general knowledge and facts. Our experiments show that *Differential Cosine Bias* measure is a potential indicator of bias in NLP models compared to the popular bias evaluation measures.

This chapter includes a substantial amount of text, figures and tables from papers whose primary author is the same as this dissertation, and published by IEEE in the following research paper:

- © 2022 IEEE. Reprinted, with permission, from Sayenju, Sudhashree, Ramazan Aygun, Bill Franks, Sereres Johnston, George Lee, Dan Sullivan, and Girish Modgil. "Stereotype and Categorical Bias Evaluation via Differential Cosine Bias Measure." In 2022 IEEE 16th International Conference on Big Data (Big Data).

## 4.1 Motivation

In the past, the success of computers depended mostly on meaningful and powerful processing of structured data. Recently, advanced technologies and research in Natural Language Processing (NLP) have enabled computers to even understand unstructured text data [133]. Despite languages composed of complex structures and full of irregularities, NLP systems can attempt to accomplish sophisticated tasks such as translation, question-answering, summarization, classification and other forms of text comprehension. Although running the most advanced language models such as BERT [134] variants or GPT-3 [12] on some sample tasks appear to mimic human response, those models are not free of biases. Since humans are affected by the outcomes of these NLP models, it is necessary to ensure fairness. The first step to fairness is to quantify undesired bias in NLP models or embeddings properly. Otherwise, not recognizing potential biases pose a harm to the society.

In this chapter, we classify biases as unfavorable stereotype biases and categorical biases. Societal stereotype biases [37] are preconceived opinions, attitudes and judgement that are either positive or negative attributes to a group of people identified by various demographics like gender, age, race, ethnicity, language, nationality, disability, sexual orientation, etc. On the other hand, categorical biases cover the general knowledge of the world. For example, an NLP model should not glean that some occupations are more appropriate for a specific gender, but it is also important for the model to learn that doctors perform surgeries, not the teachers. Similarly, it is important that the model can generate similar embeddings for 'the United States of America' and 'Canada' due to their geographical locations separated by a border. However, we do not want the model to relate 'United States of America' closely to White people but farther to people of other races.



Most of the research studies in NLP bias have solely focused on measuring and mitigating unfavorable human stereotype biases in NLP systems [38, 35, 39, 40, 34, 41, 42, 32]. Parallely, it is very important that the model learns common knowledge and facts from the text corpus displaying categorical bias that is necessary. Especially when some NLP models are built only for a specific domain, it is essential that the domain specific categorical biases are present. It should be noted that undesired societal stereotype biases need to be mitigated in NLP models but categorical biases are vital for the purpose of the NLP model application.

The data-driven models tend to inherit or even amplify the biases present in the data. Moreover, the non-interpretable nature of deep learning models might even show biases that do not exist in the data [47]. Therefore, when detecting bias it is important to look into the pre-trained word embeddings as well as the output of layers of the model. One of the most widely used technique for bias evaluation is the Word Embedding Association Test (WEAT) [35]. Originally, WEAT was used on GloVe [45] embeddings but its cosine based formula can be easily implemented on other embeddings or models. The Scoring Association Means of word Embeddings (SAME) [36] measure claims to be an improved version of the bias measure. By conducting various experiments, we found out that well-known bias evaluation measures like WEAT [35] and SAME [36] do not always measure bias accurately. Some results indicated that techniques measure the co-occurrence of words or terms that they learned from the text corpus they were trained on. Another, drawback of these measures is that they are not comparable. In other words, if the values for a bias measure of one example of gender bias is higher than that of a second example, it does not necessarily mean the first example is more biased than the second. This inconsistency questions their trustworthiness for comparing two types of different types of biases.

In this chapter, we introduce a novel bias evaluation method called *Differential Cosine Bias* measure. We have developed sixteen sets of words to understand various types of biases, where some of these sets are inspired from the literature. The major challenge for bias measures is whether their values indicate the actual bias in the models or not.

Through a series of experiments, we show that our measure is an indicator of potential bias by comparing the undesired stereotypical bias that the model captures against on categorical biases that the model should learn.

### 4.1.1 Definition of Sets

We use a similar concept of target and attribute sets like in WEAT [35]. There are two *anchor bias* sets  $X$  and  $Y$ , and two *test pair* sets are  $A$  and  $B$ . The anchor bias sets are used to represent bias by providing representative words of opposite classes (e.g., male, female). Then, our measure checks whether the difference between anchor bias sets is similar to the difference of test pair sets to determine whether a similar bias exists. Our claim is that there is bias when the difference of *anchor bias* sets embeddings and the difference of *test pair* embeddings are geometrically aligned. This would infer that anchor  $X$  is biased towards test  $A$ , while anchor  $Y$  is biased towards test  $B$ . In other words, if the difference between  $X$  and  $Y$  is aligned with the difference between  $A$  and  $B$ , then  $X$  is biased towards  $A$  rather than  $B$ .

## 4.2 Differential Cosine Bias (DiCoBi) Measure

In this section, we explain our differential cosine bias measure.

### 4.2.1 DiCoBi on Singleton Sets

For simplicity, we will explain the proposed measure for the case of where each word set is a singleton set (one word each, i.e.,  $(X = \{x\}, Y = \{y\}, A = \{a\}, \text{ and } B = \{b\})$ ). The embeddings of each word is represented by its vector notations  $\vec{x}, \vec{y}, \vec{a}$  and  $\vec{b}$ . For singleton sets case, DiCoBi measure is computed as follows:

$$DiCoBi(X, Y, A, B) = \cos(\vec{x} - \vec{y}, \vec{a} - \vec{b})$$

To demonstrate our intuition, we will use the gender bias example on occupations from the NLP bias literature. The classic example in this case is ‘man’ is to ‘doctor’ as

‘woman’ is to ‘nurse’.

If gender bias is present in occupation, it is likely that the difference vector of ‘man’ and ‘woman’ is aligned similarly to the difference vector of ‘doctor’ and ‘nurse’ ( $X = \{man\}$ ,  $Y = \{woman\}$ ,  $A = \{doctor\}$ , and  $B = \{nurse\}$ ). Therefore, the DiCoBi measure would be computed as:

$$\cos(\overrightarrow{man} - \overrightarrow{woman}, \overrightarrow{doctor} - \overrightarrow{nurse}) \quad (4.1)$$

For visualization purposes, we will demonstrate bias in 2 dimensions. In Figures 4.1, 4.2 and 4.3 we are analyzing the arrangement of difference vectors shown in black and green for  $\overrightarrow{man} - \overrightarrow{woman}$  and  $\overrightarrow{doctor} - \overrightarrow{nurse}$ , respectively. In case of the strongest bias, the values of DiCoBi will either be 1 (Fig. 4.1) or  $-1$  (Fig. 4.2), where the sign indicates the direction of bias. One way gender bias is the strongest is when the angle formed by cosine similarity of the difference vectors is  $0^\circ$ . This happens when the difference vectors are parallel in the same direction (Fig. 4.1 (a)) or in the opposite direction (Fig 4.1 (b)). Another manner in which gender bias is strongly replicated in occupation is when the angle formed by difference vectors is  $180^\circ$ . An angle of  $180^\circ$  is formed either when the black and green are parallel in opposite directions (Fig. 4.2 (a)) or they overlap each other pointing again in opposite directions (Fig. 4.2 (b)). On the other hand, if gender bias is the least present, we would expect the difference vectors to be aligned perpendicularly like in Fig. 4.3.

It should be noted that the cases described above are for extreme (1,  $-1$ ) or ideal (0) cases for presence and absence of bias. The presence of bias for the DiCoBi measure is evaluated based on its distance from 0.

## 4.2.2 General DiCoBi Measure

Here, we generalize the DiCoBi measure for multiple words in the anchor sets ( $\mathbf{X}, \mathbf{Y}$ ) and test sets ( $\mathbf{A}, \mathbf{B}$ ). Assume that the cardinality of the sets is represented as  $|\mathbf{X}| = m$ ,  $|\mathbf{Y}| = n$ ,  $|\mathbf{A}| = p$ , and  $|\mathbf{B}| = q$ . Let the sets of words be for  $\mathbf{X} = \{x_1, x_2, \dots, x_m\}$ ,  $\mathbf{Y} =$

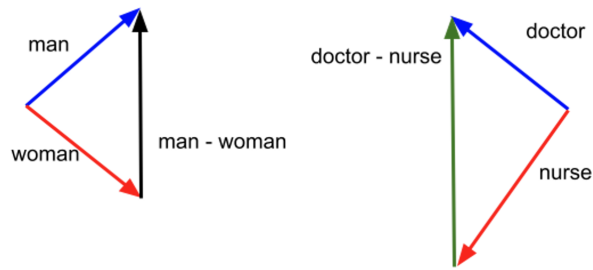


Figure 4.1: Difference vectors of gender and occupation are parallel  $\cos(\overrightarrow{man} - \overrightarrow{woman}, \overrightarrow{doctor} - \overrightarrow{nurse}) = 1$

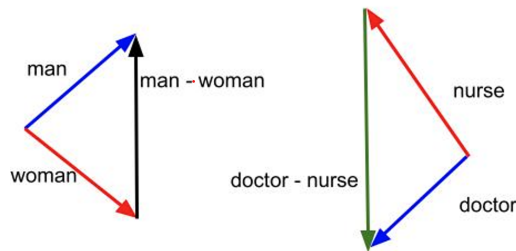


Figure 4.2: Difference vectors of gender and occupation are parallel

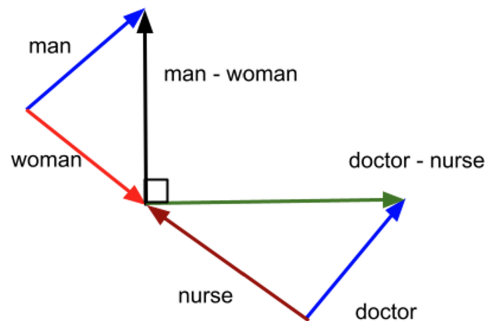


Figure 4.3: Cases where our differential cosine bias metric  $\cos(\overrightarrow{A1} - \overrightarrow{B1}, \overrightarrow{A2} - \overrightarrow{B2}) = 0$

$\{y_1, y_2, \dots, y_n\}$ ,  $\mathbf{A} = \{a_1, a_2, \dots, a_p\}$ ,  $\mathbf{B} = \{b_1, b_2, \dots, b_q\}$ . Note that the anchor bias and test pairs sets in  $\mathbf{X}$ ,  $\mathbf{Y}$ ,  $\mathbf{A}$ , and  $\mathbf{B}$  may have varying cardinalities. In other words, it is not necessary that  $m = n$  or  $p = q$ . Then,  $DiCoBi(X, Y, A, B)$  can be computed as follows:

$$DiCoBi(X, Y, A, B) = \max_{x_i \in X, y_j \in Y, a_k \in A, b_t \in B} \cos(\tilde{\mathbf{x}}_i - \tilde{\mathbf{y}}_j, \tilde{\mathbf{a}}_k - \tilde{\mathbf{b}}_t) \quad (4.2)$$

The ranges for indices  $i, j, k$ , and  $t$  are  $[1, m]$ ,  $[1, n]$ ,  $[1, p]$ , and  $[1, q]$  respectively. As this is a cosine based measure, the range of values for DiCoBi is  $[-1, 1]$ .

When we are computing DiCoBi for sets containing multiple elements, we take all possible pairs of differences  $(\mathbf{x}_i - \mathbf{y}_j, \mathbf{a}_k - \mathbf{b}_t)$  where  $\mathbf{x}_i \in \mathbf{X}$ ,  $\mathbf{y}_j \in \mathbf{Y}$ ,  $\mathbf{a}_k \in \mathbf{A}$ ,  $\mathbf{b}_t \in \mathbf{B}$ . We use the maximum (max) operator in Equation 4.2 to highlight the worst bias case in the group of words that could be of interest. We create a matrix that takes all combination of words in anchor sets  $\mathbf{X}$  and  $\mathbf{Y}$  to form the differences such that row index is from  $\mathbf{X}$  and column index is from  $\mathbf{Y}$ . Let the dimension of  $\mathbf{diff}(\mathbf{X}, \mathbf{Y})$  ( $m \times n$ ). For example, the element in the matrix denoted by  $\mathbf{diff}(\mathbf{X}, \mathbf{Y})[i, j]$  is the difference  $(x_i - y_j)$ .

Similarly, we also create a matrix that takes all combination of words in target sets  $\mathbf{A}$  and  $\mathbf{B}$  to form the differences such that row index is from  $\mathbf{A}$  and column index is from  $\mathbf{B}$ . Let us denote this matrix of differences as  $\mathbf{diff}(\mathbf{A}, \mathbf{B})$  such that its dimension is  $(p \times q)$ .

**Example using multi-element sets:** Suppose, we the following groups of words

$$X = \left[ \text{'male'} \quad \text{'man'} \right] \quad (4.3)$$

$$Y = \left[ \text{'female'} \quad \text{'woman'} \quad \text{'lady'} \right] \quad (4.4)$$

$$A = \left[ \text{'coder'} \quad \text{'scientist'} \right] \quad (4.5)$$

$$B = \left[ \text{'family'} \quad \text{'homemaker'} \right] \quad (4.6)$$

Then, the difference matrices are formed by subtracting the following combination of word embeddings:

$$\mathbf{diff}(\mathbf{X}, \mathbf{Y}) = \begin{bmatrix} ('male'-'female') & ('male'-'woman') & ('male'-'lady') \\ ('man'-'female') & ('man'-'woman') & ('man'-'lady') \end{bmatrix} \quad (4.7)$$

$$\mathbf{diff}(\mathbf{A}, \mathbf{B}) = \begin{bmatrix} ('coder'-'family') & ('coder'-'homemaker') \\ ('scientist'-'family') & ('scientist'-'homemaker') \end{bmatrix} \quad (4.8)$$

For parallelization purposes during computation, we will use  $\mathbf{diff}(\mathbf{X}, \mathbf{Y})$  and  $\mathbf{diff}(\mathbf{A}, \mathbf{B})$  to get the differential cosine similarities and filter the worst case.

When we are measuring stereotype bias, it is very important to choose the words in each set carefully. When measuring biases in groups of words, each word choice is crucial to the metric analysis. Word or sentence embeddings have been found to be very sensitive. Therefore, one should be very clear of what needs to be measured.

## 4.3 Experiments

In this section we will present experiments showing unfavorable stereotype bias on WEAT, SAME and our DiCoBi. We will first define the word groups that we use and then tabulate the results of our bias detection experiments.

### 4.3.1 Word Groups

Table 4.1 lists 16 word groups that we used in our experiments to quantify various types of biases. To check gender bias, we use Male words ( $X$ ), Female words ( $Y$ ), Career ( $A$ ) and Family ( $B$ ). To measure racial bias, we process word sets White ( $X$ ), Colored ( $Y$ ), Pleasant ( $A$ ), and Unpleasant ( $B$ ). To measure religion bias we investigate word sets Christianity ( $X$ ), Islam ( $Y$ ), Pleasant, ( $A$ ) and Unpleasant ( $B$ ). For testing the categorical biases we take Instruments ( $X$ ), Weapons ( $Y$ ), Pleasant ( $A$ ) and Unpleasant ( $B$ ). Alternatively, we take European Countries ( $X$ ), European Capitals ( $Y$ ), Asian Countries ( $A$ ), Asian Capitals ( $B$ ) in different combinations for  $\mathbf{X}, \mathbf{Y}, \mathbf{A}, \mathbf{B}$  as neutral bias should exist for these sets.

Set name	Word list
<b>Male words</b>	‘male’, ‘man’, ‘guy’, ‘father’, ‘brother’
<b>Female words</b>	‘female’, ‘woman’, ‘girl’, ‘mother’, ‘sister’
<b>Career</b>	‘professional’, ‘corporation’, ‘office’, ‘business’, ‘career’
<b>Family</b>	‘home’, ‘children’, ‘family’, ‘marriage’, ‘wedding’
<b>White</b>	‘white’, ‘caucasian’, ‘european’, ‘fair’, ‘light’, ‘blonde’, ‘brunette’
<b>Colored</b>	‘latinos’, ‘black’, ‘asian’, ‘brown’, ‘dark’, ‘middle eastern’, ‘african’
<b>Pleasant</b>	‘freedom’, ‘health’, ‘love’, ‘peace’, ‘happy’, ‘friend’, ‘heaven’
<b>Unpleasant</b>	‘abuse’, ‘crash’, ‘filth’, ‘murder’, ‘sickness’, ‘accident’, ‘death’
<b>Christianity</b>	‘Christianity’, ‘protestant’, ‘catholic’, ‘evanagelic’, ‘jesus’, ‘easter’, ‘christmas’
<b>Islam</b>	‘Islam’, ‘muslim’, ‘Sunnis’, ‘Shias’, ‘Muhammad’, ‘prophet’, ‘Quran’
<b>Instruments</b>	‘bagpipe’, ‘cello’, ‘guitar’, ‘lute’, ‘trombone’, ‘banjo’, ‘clarinet’, ‘harmonica’, ‘mandolin’, ‘trumpet’, ‘bassoon’, ‘drum’, ‘harp’, ‘oboe’, ‘tuba’, ‘bell’, ‘fiddle’, ‘harpsichord’, ‘piano’, ‘viola’, ‘bongo’, ‘flute’, ‘horn’, ‘saxophone’, ‘violin’
<b>Weapons</b>	‘arrow’, ‘club’, ‘gun’, ‘missile’, ‘spear’, ‘axe’, ‘dagger’, ‘harpoon’, ‘pistol’, ‘sword’, ‘blade’, ‘dynamite’, ‘hatchet’, ‘rifle’, ‘tank’, ‘bomb’, ‘firearm’, ‘knife’, ‘shotgun’, ‘teargas’, ‘cannon’, ‘grenade’, ‘mace’, ‘slingshot’, ‘whip’
<b>Europe</b>	‘Italy’, ‘Germany’, ‘France’, ‘Switzerland’
<b>Europe capitals</b>	‘Rome’, ‘Berlin’, ‘Paris’, ‘Bern’
<b>Asia</b>	‘Nepal’, ‘Thailand’, ‘Japan’, ‘Philippines’
<b>Asia capitals</b>	‘Kathmandu’, ‘Bangkok’, ‘Tokyo’, ‘Manilla’

Table 4.1: Word sets to test various types of biases

### 4.3.2 NLP Model for Bias Analysis

The following experiments were carried out on a pre-trained BERT (Bidirectional Encoder Representations from Transformers [134]) model namely, the version *bert-base-uncased*. Majority of research work in stereotype bias quantification use fixed word embeddings like GloVe [45] or Word2vec [43, 44]. In the case of BERT, many research studies in bias evaluation use its contextual word embeddings namely the word-piece embeddings [135]. We do not use the word-piece embeddings to convert text to its con-

textual numerical representation because the vectors were very sparse. Instead we pass the word-piece embeddings into the *bert-base-uncased* model and take the output of the last 4 layers of BERT.

### 4.3.3 Bias Validation

The biggest challenge for this problem is to determine whether there is an inherent bias in the trained model or not. Measuring bias would be meaningless if it does not reflect the actual bias in the model. From embedding vector representations, we already know that some neutral categorical bias should exist. These bias measures should indicate these categorical biases when they occur. For this purpose, Table 4.2 enlists the word level experiments to measure something neutral and obvious like a categorical bias for country and their capitals. We provide examples of correct as well as incorrect capitals to countries for pairs  $(\mathbf{X}, \mathbf{Y})$  and  $(\mathbf{A}, \mathbf{B})$ . We expect correct capital experiments to show high DiCoBi (high categorical bias) and incorrect capital examples to show low DiCoBi. When the wrong country capitals are paired, the values are low (range  $[-0.011, 0.067]$ ). Similarly, for country capitals that are correctly paired, we observe higher values for DiCoBi ranging from 0.476 to 0.742. This indicates that correct country capitals relations are similarly aligned. Thus, DiCoBi can distinguish whether or not such categorical biases are present. Nevertheless, the values for WEAT and SAME do not distinguish correct from incorrect capitals. All values for WEAT are almost 2, be it correct or incorrect capitals. The SAME scores show some range in values  $[0.009, 0.184]$  but the values cannot distinguish one group of experiments from the other. These results show that our DiCoBi measure is capable of showing the potential bias in the model.

### 4.3.4 Gender Bias

Next, we check whether gender bias exists in the NLP model or not. Table 4.3 shows that WEAT and SAME on BERT last 4 layer embeddings do not show significant difference within gender bias examples (range  $[-1.999, 2]$ ). The SAME scores show some variation giving highest gender bias for case ‘man’, ‘woman’, ‘engineer’, ‘homemaker’. DiCoBi



also reveals gender bias. The range of values for gender (stereotype) bias that is implicitly learned is low (range [-0.011, 0.198]). Our measure shows within gender bias examples, the word ‘he’ is closer to ‘doctor’ and ‘she’ is closer to ‘nurse’.

Bias type	X	Y	A	B	WEAT	SAME	DiCoBi
Correct Capital	‘Italy’	‘Rome’	‘Germany’	‘Berlin’	2.000	0.062	0.634
	‘Italy’	‘Rome’	‘France’	‘Paris’	2.000	0.063	0.621
	‘Germany’	‘Berlin’	‘France’	‘Paris’	1.999	0.053	0.742
	‘Germany’	‘Berlin’	‘Switzerland’	‘Bern’	2.000	0.044	0.522
	‘France’	‘Paris’	‘Switzerland’	‘Bern’	2.000	0.041	0.476
Incorrect Capital	‘Italy’	‘Berlin’	‘Germany’	‘Rome’	2.000	0.031	0.057
	‘France’	‘Bern’	‘Switzerland’	‘Paris’	2.000	0.168	0.067
	‘Germany’	‘Paris’	‘France’	‘Berlin’	1.999	0.046	-0.011
	‘Germany’	‘Rome’	‘Italy’	‘Berlin’	2.000	0.009	0.057
	‘France’	‘Berlin’	‘Germany’	‘Paris’	1.999	0.184	0.091

Table 4.2: Quantifying categorical biases for word level experiments

Bias type	X	Y	A	B	WEAT	SAME	DiCoBi
Gender	‘man’	‘woman’	‘doctor’	‘nurse’	2.000	0.022	0.049
	‘man’	‘woman’	‘child’	‘beautiful’	-1.999	-0.022	-0.011
	‘man’	‘woman’	‘engineer’	‘homemaker’	1.999	0.121	-0.073
	‘he’	‘she’	‘doctor’	‘nurse’	2.000	0.028	<b>0.198</b>
	‘he’	‘she’	‘child’	‘beautiful’	2.000	-0.026	0.044

Table 4.3: Quantifying gender bias for word level experiments

### 4.3.5 Word Group Level Experiments

The first three examples in Table 4.4 use word sets that are measuring various types of unfavorable stereotype biases. The first example is to measure whether Gender bias is present, i.e., male words are closer to career words and female to family. If NLP models were ideal we want them to show low bias. The second experiment is to examine whether the model has racial bias embedded in them. The third example is to investigate racial bias. The last example checks if words in Christianity associated to pleasant terms and terms related to Islam are associated to unpleasant terms.

The last three experiments in Table 4.4 are for testing categorical bias that we want the model to learn. The second word group experiment analyzes whether the differences in continents are similar in countries as they are in their capital cities. The last experiment

inspects whether the relationship between country and their capital cities are similar in different continents.

Bias type	X	Y	A	B	WEAT	Differential cosine bias
Gender	Male words	Female words	Career	Family	0.130	0.232
Race	White	Colored	Pleasant	Unpleasant	0.662	<b>0.581</b>
Religion	Christianity	Islam	Pleasant	Unpleasant	0.623	0.565
Pleasant Unpleasant	Instrument	Weapons	Pleasant	Unpleasant	0.211	0.780
Continent	Europe	Asia	Europe Capitals	Asia Capitals	1.968	0.565
Country Capital	Europe	Europe capitals	Asia	Asia Capitals	1.677	0.706

Table 4.4: Quantifying various types of biases for word group level experiments

## 4.4 Discussion

Our experiments in Tables 4.2 and 4.3 show that the well known measures like WEAT and SAME do not show proper variation in values at individual word level, be it for stereotype bias like gender or categorical bias that the model should learn as necessary knowledge to retain. For all word level experiments WEAT values are either close to -2 or 2. In Table 4.2, the third experiment for correct capitals (0.053) and last experiment in incorrect capitals (0.046) have similar SAME scores. Therefore, SAME cannot identify whether ‘Germany’ is to ‘Berlin’ as ‘France’ to ‘Paris’. We know this should not be true. DiCoBi can correctly identify that ‘Germany’ should be to ‘Berlin’ as ‘France’ to ‘Paris’ (0.742) not the vice-versa (-0.011).

Although word level embeddings can be very sensitive, DiCoBi is capable of detecting bias (Tables 4.2 and 4.3). This is one of the reasons why our measure is better over the previous bias measures like WEAT and SAME.

For the word group experiments in Table 4.4, WEAT shows an inconsistent range of values in unfavorable stereotype biases and categorical biases. We observe WEAT values for Race and Religion stereotypes is worse than pleasant/unpleasant categorical bias we expect to have. Similar to differential cosine bias, WEAT also shows worst unfavorable stereotype bias for race. However, for categorical biases WEAT is inconsistent. We expected all categorical bias values to be high. The pleasant-unpleasant bias for instrument

and weapons shows very low bias (0.211). The country capital categorical biases are very high in comparison (1.968, 1.677). The SAME score shows very low bias (range [-0.006, 0.012]) for all experiments, be it stereotype or categorical. DiCoBi is always between [-1, 1] which makes it easier to compare the degree of bias present. Our measure shows the highest degree of unfavorable stereotype bias for Race with a value of 0.581. This value is very close to that of Continent categorical bias with a value of 0.565. Therefore, we can use DiCoBi to compare implicit biases like stereotype as well as categorical biases that are expected to appear.

Observing the cases of using individual words as well as groups of words, DiCoBi gives a reliable measure of bias. The categorical bias (expected to be present) in the word and word group levels are confirmed by DiCoBi by showing high values. WEAT and SAME do not give proper variation for the different types of biases which brings question on their trustworthiness.

Our experiments were carried out on human concepts that can be explained. However, testing *DiCoBi* measure on random words not pertaining to any particular concept is yet to be explored. It is unknown to what extent DiCoBi is looking at random difference in embeddings and to what extent actual bias is present. Possible future work lies in separating the randomness or noise and actual bias in embeddings.

## 4.5 Summary

Given a myriad of NLP applications that are readily being used every second, it is unfortunate that most of these models have not been analyzed for biases before being deployed. The bias evaluation measures have been mainly designed to detect unfavorable human stereotype bias. While it is important to scan for unfavorable stereotype biases in NLP systems, it is equally important to measure categorical biases, i.e., necessary for the models to learn from the text corpus. Previously developed bias measures namely WEAT and SAME are very sensitive when measuring bias at individual word embeddings. Additionally, they are not always comparable when measuring stereotype and categorical bias. This chapter presented our *Differential Cosine Bias (DiCoBi)* measure which can be

used to measure both categorical and unfavorable stereotype biases in NLP models and embeddings. DiCoBi measure is a cosine similarity based measure whose values are in range  $[-1, 1]$ . If bias is low, the measure should yield values that are extremely close to 0. We also show via experiments that our bias evaluation measure is capable of quantifying stereotype as well as categorical bias at individual word level and word group level. Since unfavorable stereotype bias is not desired, we would want NLP models to show low values that are almost 0. On the other hand, categorical biases are desired since some factual knowledge need to be embedded in NLP models. Our experiments on different sets of words relevant to unfavorable gender, racial, or religion bias show that our DiCoBi measure is capable of measuring stereotype bias while maintaining categorical bias.

## Chapter 5

# Quantifying Domain Knowledge for Evaluating Domain Bias

Transformer based large language models such as BERT [10] have demonstrated the ability to derive contextual information from the words surrounding it. However, when these models are applied in specific domains such as medicine, insurance, or scientific disciplines, publicly available models trained on general knowledge sources such as Wikipedia, it may not be as effective in inferring the appropriate context compared to domain-specific models trained on specialized corpora. Given the limited availability of training data for specific domains, pre-trained models can be fine-tuned via transfer learning using relatively small domain-specific corpora. However, there is currently no standardized method for quantifying the effectiveness of these domain-specific models in acquiring the necessary domain knowledge. The approaches to understand domain bias could be categorized into two: evaluating the performance of trained models for domain specific tasks such as classification or the tendency of word embeddings could be analyzed if they are closer to their domain related words. In this chapter, our approach belongs to the latter category making it generalizable without being limited to varying domain specific tasks.

## 5.1 Motivation

Although most of the time the word *bias* denotes a negative implication when having high values, in the context of *domain bias*, having large values means the language model is learning the custom words, abbreviations and knowledge from the domain text corpus. Therefore, having high domain bias would be a positive indication of successful customization of models. Using domain bias measure is specifically necessary when words are polysemic such that there is one meaning in the layman sense and a completely different meaning in the domain.

In some domains, we want the meaning of some words to be adjusted as required in the context. Especially for words or acronyms that have multiple meanings, it becomes important that the model defaults to the domain specific meaning rather than the layman meaning. It should be noted that this kind of bias is desired and essential for the model to learn. Unlike stereotype bias, we want such domain biases to be strongly present in the model. Paying attention to a critical word at a single time becomes important. To address this issue, we explore hidden layer embeddings and introduce *domain gain* measure to quantify the ability of a model to infer the correct context.

## 5.2 Domain Knowledge

When fine-tuning a model for a specific domain, it is crucial to ensure that the model learns the vocabulary and associated semantics of the domain. To achieve this goal, we consider the following three sets:

$x$  = Critical *domain* word with multiple meanings

$A$  = Layman control words

$B$  = Relevant domain words

Let us consider average embedding difference of  $x$  across layman  $A$  to be  $\mu_{\|\vec{x}-\vec{a}\|}$ . Sim-

ilarly, average embedding difference of  $x$  across domain  $B$  is  $\mu_{\|\vec{x}-\vec{b}\|}$ .

for  $a \in A, b \in B$

$$\text{domain\_gain\_difference}(x, A, B) = \mu_{\|\vec{x}-\vec{a}\|} - \mu_{\|\vec{x}-\vec{b}\|}$$

$$\text{domain\_gain}(x, A, B) = \begin{cases} \frac{\mu_{\|\vec{x}-\vec{a}\|} - \mu_{\|\vec{x}-\vec{b}\|}}{\min(\mu_{\|\vec{x}-\vec{a}\|}, \mu_{\|\vec{x}-\vec{b}\|})} & \text{if } \in [-1, 1] \\ \min\left(\frac{\mu_{\|\vec{x}-\vec{a}\|} - \mu_{\|\vec{x}-\vec{b}\|}}{\min(\mu_{\|\vec{x}-\vec{a}\|}, \mu_{\|\vec{x}-\vec{b}\|})}, 1\right) & \text{if } > 1 \\ \max\left(\frac{\mu_{\|\vec{x}-\vec{a}\|} - \mu_{\|\vec{x}-\vec{b}\|}}{\min(\mu_{\|\vec{x}-\vec{a}\|}, \mu_{\|\vec{x}-\vec{b}\|})}, -1\right) & \text{if } < -1 \end{cases}$$

Table 5.1 enlists the indication for different values of  $\text{domain\_gain\_difference}(x, A, B) =$

$$\mu_{\|\vec{x}-\vec{a}\|} - \mu_{\|\vec{x}-\vec{b}\|}.$$

$\mu_{\ \vec{x}-\vec{a}\ } - \mu_{\ \vec{x}-\vec{b}\ }$	Indication
positive ( $> 0$ )	word $x$ tends towards domain knowledge ( $x$ close to $B$ )
equal ( $= 0$ )	word $x$ is equidistant to layman and domain context
negative ( $< 0$ )	word $x$ tends towards layman context ( $x$ close to $A$ )

Table 5.1: Indications for values of average difference in magnitude

### 5.3 Model Tendency Visualization

The measure  $\text{domain\_gain}$  is normalized and falls in the range  $[-1,1]$ . Normalizing will enable us to make comparisons across experiments. As shown in Figure 5.1, if the  $\text{domain\_gain}$  value is close to 1,  $x$  tends towards the domain and if the values is close to -1,  $x$  tends towards its layman meaning. The values near 0 are defined to be in the Neutral Zone. When values are very close to 0, in either the positive or negative side, they might not be significant enough to be showing layman or domain tendencies. Therefore defining a neutral zone helps to alleviate the risk of strong classification of words to lie in layman or domain zones. Moreover, neutral zone aids in the decision of whether or not bias mitigation is necessary.

The Neutral Zone can be set by boundaries  $\alpha_-$  and  $\alpha_+$ , which can be determined by an expert. It is not necessary that  $\alpha_- = -\alpha_+$ . If  $0 < \text{domain\_gain} < \alpha_+$ , we call them to

lie in the *Neutral +ve* (positive) zone. Similarly, if  $\alpha_- < domain\_gain < 0$ , we call them to lie in the *Neutral -ve* (negative) zone. We propose a very simple method to determine  $\alpha_-$  and  $\alpha_+$  in Section 5.3.1 empirically.

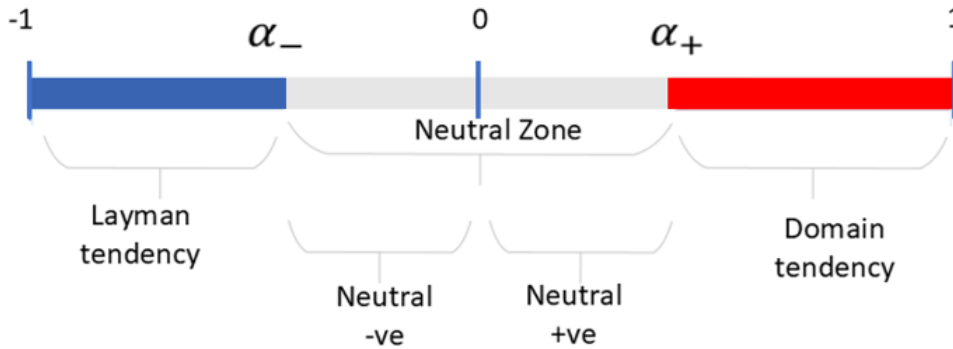


Figure 5.1: Range of *domain\_gain*

### 5.3.1 Proposed Method for Selecting $\alpha_-$ and $\alpha_+$

For simplicity, we propose to set  $\alpha_- = -\alpha_+$ . The *domain\_gain* values are designed to be compared across two or more models. Since the motive of *domain\_gain* is to check whether transfer learning is effective, we consider a pre-trained model on generic text corpus as the base model. This model is then used for one or more iterations of transfer learning on domain text corpus. Therefore, when making comparisons, the architecture of the models or the size of embeddings of the models is identical. We will use the *domain\_gain* values on the base model only to determine  $\alpha_-$  and  $\alpha_+$ . As a starting point, we analyze the experiments on the base model and then take the absolute value of *domain\_gain* and calculate the 80<sup>th</sup> percentile and consider up to its first decimal value. For example, if the 80<sup>th</sup> percentile is 0.37 we consider  $\alpha_- = -0.3$  and  $\alpha_+ = 0.3$

### 5.3.2 Arrows and their implications

We will use arrows to present how tendencies change from a base model (*Model\_General*) to a domain model (*Model\_Domain*). The tendencies are based on the boundaries set by  $\alpha_-$  and  $\alpha_+$ . Left arrows indicate negative tendency (layman tendency increases). Right



arrows indicate positive tendency (domain knowledge is gained). Ideally we want our arrows to point right as an indication of domain tendency. If an arrow has only one color, there is no change in zones. In the most extreme case, an arrow can transition two zones and have three colors. Table 5.2 lists different types of same zone arrows, their descriptions and tendencies. Table 5.3 presents a few examples of transition zone arrows but the list of arrows is not exhaustive.







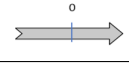


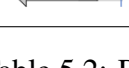
Arrow	Description	Tendency
	Domain gain further increases	Positive
	Domain gain further decreases	Negative
	Layman gain further increases	Negative
	Layman gain further decreases	Positive
	Further Neutral <sup>+</sup> increase	Positive
	Further Neutral <sup>+</sup> decrease	Negative
	Further Neutral domain gain	Positive
	Further Neutral domain loss	Negative
	Further Neutral <sup>-</sup> increase	Positive
	Further Neutral <sup>-</sup> decrease	Negative

Table 5.2: Description and tendencies of same zone arrows.



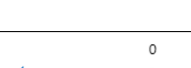
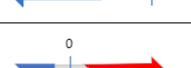
Arrow	Description	Tendency
	Word shifts from Domain to Neutral <sup>+</sup>	Negative
	Word shifts from Layman to Neutral <sup>+</sup>	Positive
	Word shifts from Neutral <sup>-</sup> to Layman	Negative
	Word shifts from Layman to Domain	Positive

Table 5.3: Examples of some transition zone arrows with their description and tendencies.

Ideally, the acquisition of domain knowledge should result in a positive *domain\_gain*.

However, it is possible for models to exhibit negative *domain\_gain* values. In such a situation, the model that consistently demonstrating higher positive values than its counterparts may be deemed to possess a greater capacity for domain knowledge acquisition and default tendencies.

## 5.4 Experiments

We carried out our experiments in a publicly available BERT model namely *bert-based-cased (Model\_General)* and a medical insurance fine-tuned model *bert-fine-tuned-medical-insurance-ner (Model\_Domain)*. In all experiments, we used the average of last 4 layers as embeddings for both models. We took some critical words ( $x$ ) that show polysemy. Set  $A$  contains words related to the layman meaning of  $x$  and set  $B$  related to the medical insurance domain meaning of  $x$ .

Table 5.4 presents the *domain\_gain* for the two models across various critical words. With our proposed method of using 80<sup>th</sup> percentile (0.166) up to first decimal on the *Model\_General*, we set the Neutral zone to be defined by boundaries  $\alpha_- = -0.1$  and  $\alpha_+ = 0.1$ . The visualization of model tendency change from *Model\_General* to *Model\_Domain* is shown in column *Tendency change*. Table 5.5 summarizes the average of the results presented in Table 5.4.

In *Model\_General* all words except for ‘examination’, ‘Resident’, and ‘Private’ exhibit a negative *domain\_gain* or a strong tendency towards layman context. Conversely, in *Model\_Domain*, the *domain\_gain* has a consistently higher positive value compared to *Model\_General* for all words except ‘Private’ (0.106 to 0.089) and ‘shot’ (-0.165 to -0.174). As a result, the majority of words in *Model\_Domain* demonstrate tendencies towards the domain. This indicates that the model has acquired domain knowledge pertaining to medical insurance. However, it should be noted that if the words ‘Private’ and ‘shot’ are crucial in the context of medical domain words ( $B$ ), then this may suggest that training on *Model\_Domain* alone may not be sufficient. In such a scenario, additional text corpora incorporating those words in their relevant domain context may be required to attain the desired default tendency of the model.

x	A (Layman control words)	B (Domain words)	domain_gain		Tendency change
			Model_General	Model_Domain	
examination	school, university, semester, grade, study, finals, mid-terms, quiz, interview	CT, CAD, X-ray, ECG, EKG, MRI, biopsy, autopsy	0.168	0.200	
Resident	inhabitant, habitant, indweller, occupant, local, citizen, tenant	doctor, hospital, health, medical, graduate, training, specialized, patients, wards, operation	0.010	0.013	
heart	love, affection, compassion, spirit	organ, blood, circulation, cardiac, vascular, artery, valve, failure	-0.005	0.013	
drug	poison, recreational, dope, opiate, narcotic, LSD, heroin, hashish, addiction, rehab	medicine, cure, pharmaceutical, remedy, health, disease, vaccine, pill, ointment	-0.041	-0.031	
Private	confidential, secret, intimate, concealed	insurance, coverage, plan, provider	0.106	0.089	
Premium	excellent, superior, prize, boon, perk, prime	insurance, price, coverage, fee, dividend, value, plan	-0.011	0.001	
admit	affirm, concede, disclose	inpatient, outpatient, beds	-0.033	-0.025	
shot	bullet, dart, missile	injection, vaccine, disease, prevention, virus, bacteria, tetanus, hepatitis	-0.165	-0.174	
blood	death, war, kinship, ancestry, lineage, family	test, culture, hemoglobin, plasma, type, RBC, WBC, tissue, platelets, fluid, arteries, veins	-0.173	-0.116	

Table 5.4: The *domain\_gain* for various tests in models *Model\_General* and *Model\_Domain*.

Our *domain\_gain* measure can be used during various training steps of an NLP model to check whether necessary knowledge of the domain is being learned.

## 5.5 Summary

We present the *domain\_gain* measure to quantify whether the default tendencies of an NLP model on a polysemic word lies towards the layman or domain meanings. In our experiments, we showed that the publicly available model *Model\_General* (*bert-base-uncased*) shows a strong tendency towards the layman meanings rather than the medical


	<b>Model_General</b>	<b>Model_Domain</b>	<b>Tendency change</b>
Average <i>domain_gain</i>	-0.016	-0.003	
Overall model summary	Domain tendency	Neutral Negative	Model_Domain is learning but its domain knowledge can still be improved. Overall we observe positive tendency. Therefore, Model_Domain is more suitable than Model_General for using in medical insurance domain.

Table 5.5: Knowledge gain summary of models

insurance context. This could be a limitation for applications relying on medical insurance domain knowledge. On the other hand, our analysis indicates that *Model\_Domain (bert-fine-tuned-medical-insurance-ner)* pulls the default tendencies of the model more towards the medical insurance domain words, thus rendering it more appropriate for use in this specific domain.

# Chapter 6

## Conclusions and Future Work

The objective of this dissertation is to quantify biases of various types. Different types of bias can be integrated into a Large Language Model (LLM) at various stages of modelling. While most of the biases such as class level and undesirable stereotype bias are unwanted, good kinds of bias such as domain bias are needed for their applications in their respective domains. To ensure fairness of these LLMs, bias quantification needs to be carried out before mitigating the unfavorable kinds of bias and strengthening the favorable types of bias.

A combination of class imbalance, semantic noise and insufficient training data tend to make LLMs favor one class more than the other. While a lot of studies have focused on stereotyping bias of humans, little work has been done on a model's class related bias. This paper introduced *directional pairwise class confusion bias* to indicate a model's favoring of a class compared to another class. We quantified and visualized this bias to reveal biased pairs. Furthermore, we also presented sample strategies to mitigate the bias using a secondary classifier. Priori bias mitigator uses a subset of the original training set for biased class pairs. The posteriori bias pair classifier uses the original training set but selects the training set based on the predictions of the original classifier. Even for cases where mitigation is limited, directional class confusion bias still gives insights about the cases that are hindering the performance of the model.

When using bias evaluation methods on LLMs for stereotype bias, it should go hand in hand with verifying it for categorical bias (facts and knowledge we expect the model

to learn). Although the most popular bias evaluation measures like WEAT and SAME claim to quantify stereotype biases, our experiments showed they are not comparable when quantifying both categorical and stereotype bias. We proposed a novel stereotype bias measure which also works for quantifying categorical bias called *Directional Cosine Bias (DiCoBi)*. As the name suggests, DiCoBi is cosine based measure, and thus, its values range is  $[-1, 1]$ . If bias is low DiCoBi will yield values close to 0. By conducting experiments at word level and word group level, we show that DiCoBi is capable of measuring stereotype bias while maintaining categorical bias.

Lastly, this dissertation proposes a novel technique to quantify domain bias. For domain use, LLMs that were trained on generic text corpus like Wikipedia are fine-tuned on domain text corpus. However, there is lack of research that quantifies acquisition of domain knowledge. We propose *domain\_gain*, as a measure to quantify whether the default tendencies of a LLM on a polysemic word lies towards the domain-related meanings or their colloquial meanings.

We proposed the following in this dissertation:

- A novel method to quantify class level bias called *directional pairwise class confusion bias*,
- A novel technique to quantify stereotype bias namely *Directional Cosine Bias measure*, and
- A novel measure to quantify the domain bias for using polysemic words, named as *domain\_gain*.

## 6.1 Future Work

As we observe exponential upscaling of every new LLM that is released, the difficulty in quantifying biases also increases at the same pace. In addition to the methods proposed in this research, new methodologies are necessary to deal with the additional sophistication

of LLMs. Additionally, the methodologies proposed in this dissertation can be further refined .

In order to mitigate our proposed class level bias (*directional pairwise class confusion bias*), we used secondary models to improve the results. In the future, one could use more advanced mitigation techniques such as loss function modification or bootstrapping techniques. Currently, *directional pairwise class confusion bias* only works for pairs of classes. A future research direction could explore how a hierarchical system of classification labels could exhibit such bias. A major contributor for *directional pairwise class confusion bias* is class imbalance distribution. Although other factors also play a role in this class level bias, finding class composition threshold that contributes to the bias might be helpful.

When quantifying stereotype bias using our *DiCoBi* measure, our experiments included gender, race and religion stereotypes. However, less popular stereotypes such as disability, ideologies and political beliefs need to be explored in the future. The challenge in quantifying stereotype bias is in working without ground truth. It is still unknown to what extent the results we obtained using *DiCoBi* showed random difference in embeddings and to what extent actual bias is present. We have used a set of control words or word pairs to determine the presence of bias. There is potential research in separating the randomness or noise and actual bias in embeddings. The next step after bias evaluation is its mitigation. Mitigation techniques should be devised before and during modelling stage instead of after its deployment.

We quantified domain bias in BERT by looking at what other sets of words it is closer to. However, we still require to explore a way to quantify bias in contextual embeddings. One could also explore ways to enhance domain bias. A possible solution would be to train the model with more domain text corpora. Another possibility is to feed the model with additional inputs where we force the polysemic word to appear with its intended domain meaning or related words multiple times in the text corpus.

# Bibliography

- [1] C Weaver Shannon and Warren Weaver. “W.:(1949) The Mathematical Theory of Communication”. In: *Press UoI, editor* (1948).
- [2] Noam Chomsky. “Logical structure in language”. In: *Journal of the American Society for Information Science* 8.4 (1957), p. 284.
- [3] Eric Brill et al. “Deducing linguistic structure from the statistics of large corpora”. In: *Proceedings of the 5th Jerusalem Conference on Information Technology, 1990. 'Next Decade in Information Technology'*. IEEE. 1990, pp. 380–389.
- [4] Mahesh V Chitrao and Ralph Grishman. “Statistical parsing of messages”. In: *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*. 1990.
- [5] Peter F Brown et al. “Word-sense disambiguation using statistical methods”. In: *29th Annual meeting of the Association for Computational Linguistics*. 1991, pp. 264–270.
- [6] Peter F. Brown et al. “A STATISTICAL APPROACH TO LANGUAGE TRANSLATION”. In: *Proceedings of the International Conference on Computational Linguistics (COLING)*. 1988.
- [7] Peter F. Brown et al. “A statistical approach to machine translation”. In: *Computational Linguistics* 16.2 (1990), pp. 76–85.
- [8] Jürgen Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural Networks* 61 (2015), pp. 85–117. issn: 0893-6080. doi: <https://doi.org/>



- 10.1016/j.neunet.2014.09.003. URL: <https://www.sciencedirect.com/science/article/pii/S0893608014002135>.
- [9] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [10] Jacob Devlin et al. “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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Tamar Solorio. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423. URL: <https://doi.org/10.18653/v1/n19-1423>.
- [11] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.
- [12] Tom B Brown et al. “Language models are few-shot learners”. In: *arXiv preprint arXiv:2005.14165* (2020).
- [13] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].
- [14] Wikipedia contributors. *Plagiarism — Wikipedia, The Free Encyclopedia*. [Online; accessed 22-December-2021]. 2004. URL: <https://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350>.
- [15] Michael V’olske et al. “TL;DR: Mining Reddit to Learn Automatic Summarization”. In: *Proceedings of the Workshop on New Frontiers in Summarization*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 59–63. DOI: 10.18653/v1/W17-4508. URL: <https://www.aclweb.org/anthology/W17-4508>.
- [16] Google Inc. *Google News corpus*. [Online; accessed 22-December-2021]. URL: <https://drive.google.com/file/d/0B7XkCwpI5KDYN1NUTTT1SS21pQmM/edit?usp=sharing>.

- [17] Tolga Bolukbasi et al. “Man is to Computer Programmer as Woman is to Home-maker? Debiasing Word Embeddings”. In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. Ed. by Daniel D. Lee et al. 2016, pp. 4349–4357. URL: <https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html>.
- [18] Kaytlin Chaloner and Alfredo 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*. 2019, pp. 25–32.
- [19] Ninareh Mehrabi et al. “A survey on bias and fairness in machine learning”. In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–35.
- [20] Ashley Z Guo et al. “Adaptive enhanced sampling by force-biasing using neural networks”. In: *The Journal of chemical physics* 148.13 (2018), p. 134108.
- [21] Anjalie Field et al. “A Survey of Race, Racism, and Anti-Racism in NLP”. In: *CoRR* abs/2106.11410 (2021). arXiv: 2106.11410. URL: <https://arxiv.org/abs/2106.11410>.
- [22] Ismael Garrido-Muñoz et al. “A Survey on Bias in Deep NLP”. In: *Applied Sciences* 11.7 (2021). ISSN: 2076-3417. DOI: 10.3390/app11073184. URL: <https://www.mdpi.com/2076-3417/11/7/3184>.
- [23] Su Lin Blodgett et al. “Language (Technology) is Power: A Critical Survey of ”Bias” in NLP”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky et al. Association for Computational Linguistics, 2020, pp. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485. URL: <https://doi.org/10.18653/v1/2020.acl-main.485>.
- [24] Anjalie Field et al. “A Survey of Race, Racism, and Anti-Racism in NLP”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational*

- Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Ed. by Chengqing Zong et al. Association for Computational Linguistics, 2021, pp. 1905–1925. doi: 10.18653/v1/2021.acl-long.149. URL: <https://doi.org/10.18653/v1/2021.acl-long.149>.
- [25] Rachel K. E. Bellamy et al. “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias”. In: *CoRR abs/1810.01943* (2018). arXiv: 1810.01943. URL: <http://arxiv.org/abs/1810.01943>.
- [26] Amazon.com LLC Amazon Web Services Inc. *AWS (Amazon Web Services) SageMaker Clarify*. <https://aws.amazon.com/sagemaker/clarify/>. Last accessed on 2021-08-30. Dec. 2020.
- [27] Haibo He and Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications*. Wiley-IEEE Press, 2013.
- [28] Nathalie Japkowicz and Shaju Stephen. “The class imbalance problem: A systematic study”. In: *Intelligent data analysis 6.5* (2002), pp. 429–449.
- [29] Justin M Johnson and Taghi M Khoshgoftaar. “Survey on deep learning with class imbalance”. In: *Journal of Big Data 6.1* (2019), pp. 1–54.
- [30] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. “A systematic study of the class imbalance problem in convolutional neural networks”. In: *Neural Networks 106* (2018), pp. 249–259.
- [31] Jesse Vig. “A Multiscale Visualization of Attention in the Transformer Model”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2019, pp. 37–42.
- [32] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. “Investigating gender bias in bert”. In: *Cognitive Computation 13.4* (2021), pp. 1008–1018.

- [33] Chandler May et al. “On Measuring Social Biases in Sentence Encoders”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 622–628. doi: 10.18653/v1/n19-1063. URL: <https://doi.org/10.18653/v1/n19-1063>.
- [34] Yi Chern Tan and L Elisa Celis. “Assessing social and intersectional biases in contextualized word representations”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [35] Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. “Semantics derived automatically from language corpora contain human-like biases”. In: *CoRR* abs/1608.07187 (2016). arXiv: 1608.07187. URL: <http://arxiv.org/abs/1608.07187>.
- [36] Sarah Schröder et al. “The SAME score: Improved cosine based bias score for word embeddings”. In: *arXiv preprint arXiv:2203.14603* (2022).
- [37] Martie G Haselton, Daniel Nettle, and Damian R Murray. “The evolution of cognitive bias”. In: *The handbook of evolutionary psychology* (2015), pp. 1–20.
- [38] Tolga Bolukbasi et al. “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: <https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f31609456-Paper.pdf>.
- [39] Moin Nadeem, Anna Bethke, and Siva Reddy. “Stereoset: Measuring stereotypical bias in pretrained language models”. In: *arXiv preprint arXiv:2004.09456* (2020).

- [40] Malvina Nissim, Rik van Noord, and Rob van der Goot. “Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor”. In: *CoRR* abs/1905.09866 (2019). arXiv: 1905.09866. URL: <http://arxiv.org/abs/1905.09866>.
- [41] Yupei Du, Qixiang Fang, and Dong Nguyen. “Assessing the Reliability of Word Embedding Gender Bias Measures”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 10012–10034.
- [42] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. “Understanding Undesirable Word Embedding Associations”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1696–1705. DOI: 10.18653/v1/P19-1166. URL: <https://aclanthology.org/P19-1166>.
- [43] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems* 26 (2013), pp. 3111–3119.
- [44] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).
- [45] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. “GloVe: Global Vectors for Word Representation”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: <http://www.aclweb.org/anthology/D14-1162>.
- [46] Matthew E Peters et al. *Deep contextualized word representations*. *NAACL 2018*.
- [47] Harini Suresh and John V. Gutttag. “A Framework for Understanding Unintended Consequences of Machine Learning”. In: *CoRR* abs/1901.10002 (2019). arXiv: 1901.10002. URL: <http://arxiv.org/abs/1901.10002>.
- [48] Deven Shah, H Andrew Schwartz, and Dirk Hovy. “Predictive biases in natural language processing models: A conceptual framework and overview”. In: *arXiv preprint arXiv:1912.11078* (2019).

- [49] Lucas Dixon et al. “Measuring and mitigating unintended bias in text classification”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 67–73.
- [50] Su Lin Blodgett et al. “Language (Technology) is power: A critical survey of ”Bias” in NLP”. In: *arXiv preprint arXiv:2005.14050* (2020).
- [51] Ben Green. “Good” isn’t good enough”. In: *Proceedings of the AI for Social Good workshop at NeurIPS*. 2019.
- [52] Ismael Garrido-Muñoz et al. “A Survey on Bias in Deep NLP”. In: *Applied Sciences* 11.7 (2021), p. 3184.
- [53] Tony Sun et al. “Mitigating gender bias in natural language processing: Literature review”. In: *arXiv preprint arXiv:1906.08976* (2019).
- [54] Tolga Bolukbasi et al. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: *Advances in neural information processing systems* 29 (2016), pp. 4349–4357.
- [55] Jieyu Zhao et al. “Men also like shopping: Reducing gender bias amplification using corpus-level constraints”. In: *arXiv preprint arXiv:1707.09457* (2017).
- [56] Jieyu Zhao et al. “Gender bias in coreference resolution: Evaluation and debiasing methods”. In: *arXiv preprint arXiv:1804.06876* (2018).
- [57] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. “Mitigating unwanted biases with adversarial learning”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 335–340.
- [58] Yulia Tsvetkov, Julia Mendelsohn, and Dan Jurafsky. “A framework for the computational linguistic analysis of dehumanization”. In: *Frontiers in artificial intelligence* (2020).
- [59] Ben Hutchinson et al. “Social biases in NLP models as barriers for persons with disabilities”. In: *arXiv preprint arXiv:2005.00813* (2020).

- [60] Anjalie Field et al. “A survey of race, racism, and anti-racism in NLP”. In: *arXiv preprint arXiv:2106.11410* (2021).
- [61] Su Lin Blodgett, Johnny Wei, and Brendan O’Connor. “Twitter universal dependency parsing for African-American and mainstream American English”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 1415–1425.
- [62] Zeerak Waseem. “Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter”. In: *Proceedings of the first workshop on NLP and computational social science*. 2016, pp. 138–142.
- [63] Pia Sommerauer and Antske Fokkens. “Conceptual change and distributional semantic models: an exploratory study on pitfalls and possibilities”. In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. 2019, pp. 223–233.
- [64] Svetlana Kiritchenko and Saif M Mohammad. “Examining gender and race bias in two hundred sentiment analysis systems”. In: *arXiv preprint arXiv:1805.04508* (2018).
- [65] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. “Racial bias in hate speech and abusive language detection datasets”. In: *arXiv preprint arXiv:1905.12516* (2019).
- [66] Tomas Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”. In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2013. URL: <http://arxiv.org/abs/1301.3781>.
- [67] Jieyu Zhao et al. “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Vol-*

- ume 2 (Short Papers)*. Ed. by Marilyn A. Walker, Heng Ji, and Amanda Stent. Association for Computational Linguistics, 2018, pp. 15–20. doi: 10.18653/v1/n18-2003. URL: <https://doi.org/10.18653/v1/n18-2003>.
- [68] Jieyu Zhao et al. “Learning Gender-Neutral Word Embeddings”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Ed. by Ellen Riloff et al. Association for Computational Linguistics, 2018, pp. 4847–4853. URL: <https://aclanthology.org/D18-1521/>.
- [69] Thomas Manzini et al. “Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Tamar Solorio. Association for Computational Linguistics, 2019, pp. 615–621. doi: 10.18653/v1/n19-1062. URL: <https://doi.org/10.18653/v1/n19-1062>.
- [70] Marzieh Babaeianjelodar et al. “Quantifying Gender Bias in Different Corpora”. In: *Companion of The 2020 Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*. Ed. by Amal El Fallah Seghrouchni et al. ACM, 2020, pp. 752–759. doi: 10.1145/3366424.3383559. URL: <https://doi.org/10.1145/3366424.3383559>.
- [71] Ben Hutchinson et al. “Social Biases in NLP Models as Barriers for Persons with Disabilities”. In: *CoRR abs/2005.00813 (2020)*. arXiv: 2005.00813. URL: <https://arxiv.org/abs/2005.00813>.
- [72] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. “Investigating Gender Bias in BERT”. In: *Cogn. Comput.* 13.4 (2021), pp. 1008–1018. doi: 10.1007/s12559-021-09881-2. URL: <https://doi.org/10.1007/s12559-021-09881-2>.



- [73] Jesse Vig. “A Multiscale Visualization of Attention in the Transformer Model”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*. Ed. by Marta R. Costa-jussà and Enrique Alfonseca. Association for Computational Linguistics, 2019, pp. 37–42. doi: 10.18653/v1/p19-3007. URL: <https://doi.org/10.18653/v1/p19-3007>.
- [74] Luciano Floridi and Massimo Chiriatti. “GPT-3: Its Nature, Scope, Limits, and Consequences”. In: *Minds Mach.* 30.4 (2020), pp. 681–694. doi: 10.1007/s11023-020-09548-1. URL: <https://doi.org/10.1007/s11023-020-09548-1>.
- [75] Tomas Mikolov et al. *word2vec*. URL: <https://code.google.com/archive/p/word2vec/>.
- [76] Thomas Manzini et al. “Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings”. In: *arXiv preprint arXiv:1904.04047* (2019).
- [77] Nathaniel Swinger et al. “What are the biases in my word embedding?” In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 305–311.
- [78] Capitolina Díaz Martínez, Pablo Díaz García, and Pablo Navarro Sustaeta. “Hidden Gender Bias in Big Data as Revealed Through Neural Networks: Man is to Woman as Work is to Mother?” In: *Revista Española de Investigaciones Sociológicas (REIS)* 172.172 (2020), pp. 41–76.
- [79] Lu Cheng, Suyu Ge, and Huan Liu. “Toward Understanding Bias Correlations for Mitigation in NLP”. In: *arXiv preprint arXiv:2205.12391* (2022). doi: 10.48550/ARXIV.2205.12391. URL: <https://arxiv.org/abs/2205.12391>.
- [80] Georgina Curto et al. “Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings”. In: *AI & society* (2022), pp. 1–16.

- [81] Yan Chen et al. “Gender Bias and Under-Representation in Natural Language Processing Across Human Languages”. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 24–34. ISBN: 9781450384735. DOI: 10.1145/3461702.3462530. URL: <https://doi.org/10.1145/3461702.3462530>.
- [82] Xiuying Chen et al. “Unsupervised Mitigating Gender Bias by Character Components: A Case Study of Chinese Word Embedding”. In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Seattle, Washington: Association for Computational Linguistics, July 2022, pp. 121–128. URL: <https://aclanthology.org/2022.gebnlp-1.14>.
- [83] Jieyu Zhao et al. “Learning Gender-Neutral Word Embeddings”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, pp. 4847–4853.
- [84] Joel Escudé Font and Marta R Costa-jussà. “Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. 2019, pp. 147–154.
- [85] Anne Lauscher et al. “A general framework for implicit and explicit debiasing of distributional word vector spaces”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34(05) (2020), pp. 8131–8138.
- [86] Sunipa Dev et al. “OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings”. In: *arXiv preprint arXiv:2007.00049* (2020).
- [87] Sunipa Dev and Jeff Phillips. “Attenuating bias in word vectors”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 879–887.
- [88] Anne Lauscher and Goran Glavaš. “Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors”. In: *Proceedings of the*

- Eighth Joint Conference on Lexical and Computational Semantics (\* SEM 2019)*. 2019, pp. 85–91.
- [89] Wei Guo and Aylin Caliskan. “Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases”. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 122–133.
- [90] Jieyu Zhao et al. “Gender Bias in Contextualized Word Embeddings”. 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)*. 2019, pp. 629–634.
- [91] Sanjana Marcé and Adam Poliak. “On Gender Biases in Offensive Language Classification Models”. In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Seattle, Washington: Association for Computational Linguistics, July 2022, pp. 174–183. URL: <https://aclanthology.org/2022.gebnlp-1.19>.
- [92] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. “The united nations parallel corpus v1. 0”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. 2016, pp. 3530–3534.
- [93] Philipp Koehn et al. “Europarl: A parallel corpus for statistical machine translation”. In: *MT summit*. Vol. 5. Citeseer. 2005, pp. 79–86.
- [94] Kishore Papineni et al. “Bleu: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.
- [95] Rowan Hall Maudslay et al. “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)*. 2019, pp. 5267–5275.

- [96] Anne Lauscher et al. “AraWEAT: Multidimensional Analysis of Biases in Arabic Word Embeddings”. In: *Proceedings of the Fifth Arabic Natural Language Processing Workshop*. 2020, pp. 192–199.
- [97] Kaiji Lu et al. “Gender bias in neural natural language processing”. In: *Logic, Language, and Security*. Springer, 2020, pp. 189–202.
- [98] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.
- [99] Matt Gardner et al. “Allennlp: A deep semantic natural language processing platform”. In: *arXiv preprint arXiv:1803.07640* (2018).
- [100] Ciprian Chelba et al. “One billion word benchmark for measuring progress in statistical language modeling”. In: *arXiv preprint arXiv:1312.3005* (2013).
- [101] Christine Basta, Marta R Costa-jussà, and Noe Casas. “Evaluating the Underlying Gender Bias in Contextualized Word Embeddings”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. 2019, pp. 33–39.
- [102] Alec Radford et al. “Improving language understanding by generative pre-training”. In: *OpenAI* (2018). URL: [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf).
- [103] Marion Bartl, Malvina Nissim, and Albert Gatt. “Unmasking Contextual Stereotypes: Measuring and Mitigating BERT’s Gender Bias”. In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. 2020, pp. 1–16.
- [104] Emily Sheng et al. “The Woman Worked as a Babysitter: On Biases in Language Generation”. 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)*. 2019, pp. 3407–3412.
- [105] Marzieh Babaeianjelodar et al. “Quantifying gender bias in different corpora”. In: *Companion Proceedings of the Web Conference 2020*. 2020, pp. 752–759.

- [106] Keita Kurita et al. “Measuring Bias in Contextualized Word Representations”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 166–172. DOI: 10.18653/v1/W19-3823. URL: <https://aclanthology.org/W19-3823>.
- [107] Brienna Herold, James Waller, and Raja Kushalnagar. “Applying the Stereotype Content Model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies”. In: *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 58–65. DOI: 10.18653/v1/2022.slpat-1.8. URL: <https://aclanthology.org/2022.slpat-1.8>.
- [108] Sophie Jentsch and Cigdem Turan. “Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task”. In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Seattle, Washington: Association for Computational Linguistics, July 2022, pp. 184–199. URL: <https://aclanthology.org/2022.gebnlp-1.20>.
- [109] Virginia K Felkner et al. “Towards WinoQueer: Developing a Benchmark for Anti-Queer Bias in Large Language Models”. In: *arXiv e-prints* (2022), arXiv:2206.2206.
- [110] OpenAI. *ChatGPT*. Last accessed on 2023-04-15. 2023.
- [111] Xiangyu Peng et al. “Reducing Non-Normative Text Generation from Language Models”. In: *International Conference on Natural Language Generation*. 2020.
- [112] Sophie Groenwold et al. “Investigating African-American Vernacular English in Transformer-Based Text Generation”. In: *arXiv preprint arXiv:2010.02510* (2020).
- [113] Samhita Honnavalli et al. “Towards Understanding Gender-Seniority Compound Bias in Natural Language Generation”. In: *arXiv preprint arXiv:2205.09830* (2022).

- [114] Kris McGuffie and Alex Newhouse. “The radicalization risks of GPT-3 and advanced neural language models”. In: *arXiv preprint arXiv:2009.06807* (2020).
- [115] Luciano Floridi and Massimo Chiriatti. “GPT-3: Its nature, scope, limits, and consequences”. In: *Minds and Machines* 30 (2020), pp. 681–694.
- [116] Abubakar Abid, Maheen Farooqi, and James Zou. “Persistent anti-muslim bias in large language models”. In: *arXiv preprint arXiv:2101.05783* (2021).
- [117] Google. URL: <https://www.google.com/>.
- [118] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. “Measuring individual differences in implicit cognition: the implicit association test.” In: *Journal of personality and social psychology* 74.6 (1998), p. 1464.
- [119] Chandler May et al. “On measuring social biases in sentence encoders”. In: *arXiv preprint arXiv:1903.10561* (2019).
- [120] Hila Gonen and Yoav Goldberg. “Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them”. In: *arXiv preprint arXiv:1903.03862* (2019).
- [121] Pei Zhou et al. “Analyzing and Mitigating Gender Bias in Languages with Grammatical Gender and Bilingual Word Embeddings”. In: *ACL: Montréal, QC, Canada* (2019).
- [122] Sophie Jentsch et al. “Semantics derived automatically from language corpora contain human-like moral choices”. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 37–44.
- [123] Syed Meesam Raza Naqvi et al. “Leveraging Free-Form Text in Maintenance Logs Through BERT Transfer Learning”. In: *Progresses in Artificial Intelligence & Robotics: Algorithms & Applications*. Ed. by Luigi Troiano et al. Cham: Springer International Publishing, 2022, pp. 63–75. ISBN: 978-3-030-98531-8.
- [124] Yifan Peng, Shankai Yan, and Zhiyong Lu. *Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets*. 2019. arXiv: 1906.05474 [cs.CL].

- [125] Rukhma Qasim et al. “A fine-tuned BERT-based transfer learning approach for text classification”. In: *Journal of healthcare engineering* 2022 (2022).
- [126] Min Kang, Kye Hwa Lee, and Youngho Lee. “Filtered BERT: Similarity Filter-Based Augmentation with Bidirectional Transfer Learning for Protected Health Information Prediction in Clinical Documents”. In: *Applied Sciences* 11.8 (2021). ISSN: 2076-3417. DOI: 10.3390/app11083668. URL: <https://www.mdpi.com/2076-3417/11/8/3668>.
- [127] Angie Boggust, Brandon Carter, and Arvind Satyanarayan. “Embedding comparator: Visualizing differences in global structure and local neighborhoods via small multiples”. In: *27th international conference on intelligent user interfaces*. 2022, pp. 746–766.
- [128] Haoran Xu and Philipp Koehn. “Cross-lingual bert contextual embedding space mapping with isotropic and isometric conditions”. In: *arXiv preprint arXiv:2107.09186* (2021).
- [129] Gregor Wiedemann et al. “Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings”. In: *arXiv preprint arXiv:1909.10430* (2019).
- [130] Mohammad Nuruzzaman and Omar Khadeer Hussain. “A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks”. In: *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*. 2018, pp. 54–61. DOI: 10.1109/ICEBE.2018.00019.
- [131] Sandeep A Thorat and Vishakha Jadhav. “A review on implementation issues of rule-based chatbot systems”. In: *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*. 2020.
- [132] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. “What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study”. In: *Internet Science*. Ed. by Svetlana S. Bodrunova. Cham: Springer International Publishing, 2018, pp. 194–208. ISBN: 978-3-030-01437-7.

- [133] Phillip Nelson, Namratha V. Urs, and Taraka Rama Kasichayanula. “Progress in Natural Language Processing and Language Understanding”. In: *Bridging Human Intelligence and Artificial Intelligence*. Cham: Springer International Publishing, 2022, pp. 83–103. ISBN: 978-3-030-84729-6. DOI: 10.1007/978-3-030-84729-6\_6. URL: [https://doi.org/10.1007/978-3-030-84729-6\\_6](https://doi.org/10.1007/978-3-030-84729-6_6).
- [134] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of NAACL-HLT*. 2019, pp. 4171–4186.
- [135] Yonghui Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *CoRR* abs/1609.08144 (2016). arXiv: 1609.08144. URL: <http://arxiv.org/abs/1609.08144>.