CONCEPTUALIZING SOCIAL HARMS ARISING FROM BIAS AND DISCRIMINATION IN NATURAL LANGUAGE PROCESSING: RACE, GENDER & LANGUAGE

By

Jamell Dacon

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science —Doctor of Philosophy

2023

ABSTRACT

Natural language processing (NLP) is a subfield of artificial intelligence (AI) and has become increasingly prominent in our everyday lives. NLP systems are now ubiquitous as they are capable of identifying offensive and abusive conversational content and hate speech detection on social media platforms, voice and speech recognition and transcription, news recommendation, dialogue systems and digital assistants, language generation, etc. Yet, the benefits of these language technologies do not accrue evenly to all of its users leading to harmful social impacts as NLP systems reproduce stereotypes or fallacious results. Most AI systems and algorithms are data driven and require natural language data upon which to be trained. Thus, data is tightly associated to the functionality of these algorithms and systems. These systems generate complex social implications i.e., displaying human-like social biases (e.g. gender bias) that induce technological marginalization and increased feelings of disenfranchisement.

Throughout this dissertation, I argue that how harms arise in NLP systems and who is harmed by these biases, can only be conceptualized and understood at the intersection of NLP, justice and equity (e.g., Data Science for Social Good), and the coupled relationships between language and both social and racial hierarchies. I propose to address three questions at this intersection: (1) *How can we conceptualize and quantify such aforementioned harms?*; (2) *How can we introduce a set of measurements to understand "bias" in NLP systems*; and (3) *How can we quantitatively and qualitatively ensure "fairness" in NLP systems?*.

To address these pertinent question, we attempt differentiate the two consequences of predictive bias in NLP: (1) outcome disparities (i.e., racial bias) and (2) error disparities (i.e., poor system performance) to explicate the importance of modeling social factors of language by exploiting NLP tools to examine predictive biases of both binary gender-specific (male and female) and LGBTQIA2S+ representations, and on an English language variety, i.e., African American English (AAE). Language reflects society, ideology, cultural identity, and customs of communicators, as well as their values. Therefore, natural language data, culture and systems are intertwined with social norms.

Nevertheless, social media and online services contain rich textual information on topics surrounding ethnicity, gender identity and sexual orientation–members of the LGBTQIA2S+ community and language (e.g., AAE). This facilitates the collection of large-scale corpora to study social biases in NLP systems in hopes of reducing stigmatization, marginalization, mischaracterization, or erasure of dialectal languages and its speakers, pushing back against potentially discriminatory practices (in many cases—discriminatory through oversight more than malice). In this dissertation, I propose several studies to minimize the gaps between gender, race and NLP systems' performance within the scope of the three aforementioned questions.

To my family, friends and loved ones for their support, kindness, prayers, and encouragement.

ACKNOWLEDGMENTS

During my collegiate journey, I have received invaluable help, advice, support and guidance from a multitude of amazing people.

First and foremost, I would like to thank God for allowing me to push through and attain such as degree; there were many low moments but I was able to stand again after each time I was knocked down. Next, I would like to thank my primary advisor, Dr. Jiliang Tang, for his patience, guidance, support and encouragement to continue to purse my own interest in research until I found my niche. I would also like to thank Mr. Steven Thomas, for numerous opportunities for growth, value, kindness, optimism. He has taught me that no matter how tough things may be, there is always a solution. With Dr. Tang's and Mr. Thomas' help I have achieved much more than I have imagined. I would like to extend my gratitude to my Ph.D. committee members: Dr. Hui Liu, Dr. Pan-Ning Tang and Dr. Tai-Quan Peng for all of their insightful questions and comments, support, encouragement and helpful suggestions.

In addition, I would like to thank the members of the Data Science and Engineering (DSE) Lab and the Shiu Lab at MSU. Special thanks goes out to Tyler Derr, Haochen Liu, Harry Shomer, Kenia Segura Abá, Brianna Brown, Thilanka Ranaweera, Huan Chen, Serena Lotreck, Dr. Jyothi Kumar, Dr. Melissa Lehti-Shiu who were very supportive, and Dr. Shinhan Shiu for being his intelligent enthusiastic self, bringing constant joy to his lab members while directing the Shiu Lab.

Finally, I would like to express my deepest thanks and gratitude to my dearest wife, Shaylynn Crum-Dacon, and my wonderful grandmother, Catherine Branker, as well as supportive family, friends, and colleagues for their love, encouragement and prayers during this time.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1
1.1 Motivation 1
1.2 Dissertation Contributions 3

CHAPTER 2 BIAS DETECTION IN DIALOGUE GENERATION 6
2.1 Introduction 6
2.2 Fairness Analysis in Dialogue Systems 8
2.2.1 Fairness in Dialogue systems 8
2.2.2 Parallel Context Data Construction 9
2.2.3 Fairness Measurements 10
2.2.3.1 Diversity 10
2.2.3.2 Politeness 10
2.2.3.3 Sentiment 10
2.2.3.4 Attribute Words 11
2.3 Experiment 12
2.3.1 Dialogue Models 12
2.3.1.1 The Seq2Seq Generative Model 12
2.3.1.2 The Transformer Retrieval Model 12
2.3.2 Experimental Settings 14
2.3.3 Experimental Results 14
2.4 Related Work 15
2.5 Conclusion 17

CHAPTER 3 DETECTING AND EXAMINING GENDER BIAS IN THE NEWS 18
3.1 Introduction 18
3.2 Related Works 21
3.3 Datasets 21
3.4 Bias in Gender Distribution 23
3.4.1 Gender Distribution 23
3.4.2 Experiment 23
3.5 Bias in Content 25
3.5.1 Attribute Words 25
3.5.2 Experiment 25
3.6 Bias in Wording 26
3.6.1 Sentiment Analysis 26
3.6.2 Centering Resonance Analysis 28
3.6.3 Experiment 28
3.7 Conclusion 30

CHAPTER 4 DETECTING HARMFUL ONLINE CONVERSATIONAL CONTENT TOWARDS LGBTQIA2S+ INDIVIDUALS 32
4.1 Introduction 32
4.2 Preliminaries 34

4.2.1 Problem Statement . . . 35
4.2.2 Dataset . . . 35
4.2.3 Annotation . . . 36
4.2.4 Human Evaluation . . . 38

CHAPTER 5 A MULTI-LAYERED LANGUAGE ANALYSIS: A CASE STUDY OF AFRICAN-AMERICAN ENGLISH . . . 39
5.1 Introduction . . . 39
5.2 Related Work . . . 41
5.3 Dataset and Annotation . . . 42
5.3.1 Dataset . . . 42
5.3.2 Preprocessing . . . 43
5.3.3 Annotation . . . 44
5.3.4 Human Evaluation . . . 44
5.4 Methodology . . . 45
5.4.1 Part-of-Speech (POS) Tagging . . . 45
5.4.2 Models . . . 45
5.5 Operationalization of AAE as an English Language Variety . . . 46
5.6 Conclusion . . . 46
5.7 Limitations And Ethical Considerations . . . 47

CHAPTER 6 DETECTING AND MITIGATING INHERENT LINGUISTIC BIAS IN LARGE LANGUAGE MODELS . . . 48
6.1 Introduction . . . 48
6.2 Preliminaries . . . 50
6.2.1 Problem Statement . . . 50
6.2.2 Dataset . . . 51
6.2.2.1 SNLI corpus . . . 51
6.2.2.2 MNLI corpus . . . 52
6.3 CODESWITCH Creation . . . 52
6.3.1 Data Collection . . . 52
6.3.2 Candidate Retrieval . . . 53
6.3.3 Human Evaluation . . . 54
6.4 Empirical Study and Analysis . . . 55
6.5 Debiasing Methods . . . 56
6.5.1 Counterpart Data Augmentation . . . 56
6.5.2 Language Style Disentanglement . . . 56
6.5.2.1 The LSD Framework . . . 57
6.5.2.2 An Optimization Method . . . 58
6.5.3 Experimental results . . . 59
6.6 Related Work . . . 60
6.7 Conclusion and Future Works . . . 61
6.8 Limitations And Ethical Considerations . . . 62

CHAPTER 7 CONCLUSION . . . 64

7.1 Dissertation Summary 64
7.2 Future Work 65
7.3 Concluding Remarks 66

BIBLIOGRAPHY 67

APPENDIX A BIAS DETECTION IN DIALOGUE GENERATION 82

APPENDIX B DETECTING AND EXAMINING GENDER BIAS IN THE NEWS . . . 86

APPENDIX C DETECTING HARMFUL ONLINE CONVERSATIONAL CONTENT TOWARDS LGBTQIA2S+ INDIVIDUALS 91

APPENDIX D A MULTI-LAYERED LANGUAGE ANALYSIS: A CASE STUDY OF AFRICAN-AMERICAN ENGLISH 100

APPENDIX E DETECTING AND MITIGATING INHERENT LINGUISTIC BIAS IN LARGE LANGUAGE MODELS 103

CHAPTER 1

INTRODUCTION

Natural language processing (NLP) is a subfield of artificial intelligence (AI), computer science, and linguistics focused on making human communication, such as speech and text, comprehensible to computers; NLP is used in a wide variety of everyday products and services. Some of the most common ways NLP is used are through voice-activated digital assistants on smartphones, email-scanning programs used to identify spam, and translation apps that translate a multitude of languages, and thus, has become increasingly prominent in our every day lives. NLP systems are now ubiquitous in both academia and industry as they are capable of identifying offensive and abusive conversational content and hate speech detection on social media platforms [120, 36, 142], voice and speech recognition and transcription [45, 77], news recommendation [34], dialogue systems and digital assistants [85], language generation [57], etc. Yet, the benefits of these language technologies do not accrue evenly to all of its users leading to harmful societal impacts as NLP systems reproduce gender and racial stereotypes [18, 24].

1.1 Motivation

Most AI systems and algorithms are data driven and require natural language data upon which to be trained. Thus, data is tightly associated to the functionality of these algorithms and systems. These systems generate complex social implications i.e., displaying human-like social biases (e.g. gender bias) that induce technological marginalization and increased feelings of disenfranchisement. As these systems aim to learn from natural language data, sentence and word embeddings–for example, are popular NLP tools that capture the semantic similarities of sentences and words which display human-like social biases. To consider both social and racial hierarchies sustained or intensified by current NLP computational techniques and facilitate Fairness, Accountability, Transparency and Ethics (FATE) in AI, ML, and NLP, to tackle bias and fairness issues we shift towards a human-in-the-loop paradigm to address issues surrounding gender and racial biases present in AI spaces. Moreover, by drawing on interdisciplinary fields such as a sociology, political science,

sociolinguistics, education, anthropology, psychology, and thorough engagement with relevant literature outside of NLP we aim to gain a deeper recognition of the coupled relationships between language, and racial and social hierarchies–a necessary step towards establishing a trustworthy path forward.

In this thesis, we argue that how harms arise in NLP systems and who is harmed by these biases, can only be conceptualized and understood at the intersection of NLP, fairness, social justice, diversity and equity (e.g. Data Science for Social Good), and the coupled relationships between language and both social and racial hierarchies. We propose to address three questions at this intersection:

1. *How can we conceptualize and quantify such aforementioned harms?;*
2. *How can we introduce a set of measurements to understand "bias" in NLP systems?;* and
3. *How can we quantitatively and qualitatively ensure "fairness" in NLP systems?*

To address these pertinent question, we attempt differentiate the two consequences of predictive bias in NLP: (1) ***outcome disparities*** (i.e., racial bias) and (2) ***error disparities*** (i.e., poor system performance) to explicate the importance of modeling social factors of language by exploiting NLP tools to examine predictive biases of both binary (male and female) and LGBTQIA2S+ representations, and on an English language variety, African American Language (AAE)[1]. Although AAE is spoken by millions of people across the United States, this dialect continuum is perceived to be "bad english" despite numerous studies by socio/raciolinguists and dialectologists in their attempts to quantify AAE as a legitimized language [6, 48, 11, 79]. As a consequence, conversational platforms struggle to effectively facilitate less-represented dialects and English language varieties. Language reflects society, ideology, cultural identity, and customs of communicators, as well as their values. Therefore natural language data, culture and systems are intertwined with social norms.

[1]A dialectal continuum previously known as Northern Negro English, Black English Vernacular (BEV), Black English, African American Vernacular English (AAVE), African American Language (AAL), Ebonics, and Non-standard English [79, 4, 56, 55, 6, 11, 76]. It is often referred to as African American Language (AAL) and African American English (AAE). In this work, we use the denotation AAE.

> "*[T]he common misconception [is] that language use has primarily to do with words and what they mean. It doesn't. It has primarily to do with people and what* ***they*** *mean.*" – [30]

Nevertheless, social media and online services contain rich textual information on topics surrounding ethnicity, gender identity, sexual orientation and AAE, enabling the collection of large-scale corpora to study societal biases in NLP systems in hopes of reducing stigmatization, marginalization, mischaracterization, or erasure of AAE and its speakers, pushing back against potentially discriminatory practices (in many cases, discriminatory through oversight more than malice).

Throughout this thesis, we propose several studies to minimize the gaps between gender, race and NLP systems' performance within the scope of the three aforementioned questions. In order to enable in-depth conversations about what kinds of system behaviors are harmful, in what ways, to whom, and why; we will allude to three case studies, (1) *Gender, Race, Language and Social Justice*, (2) *Gender and Sexual Identities, Orientations and Expressions*, and (3) *Language, Race and Culture* referencing several published works accepted to top-tier conferences that engage with social factors of language, affected communities and NLP systems.

1.2 Dissertation Contributions

We summarize the major contributions of this dissertation in 3-fold case studies as follows:

- We conduct a pioneering case study about the fairness issues concerning (1) *Gender, Race, Language and Social Justice*, (2) *Gender and Sexual Identities, Orientations and Expressions*, and (3) *Language, Race and Culture*
- In Chapter 2, we address the case study Gender, Race, Language and Social Justice.
  - We define the fairness in dialogue systems formally and introduce a set of measurements to understand the fairness of a dialogue system quantitatively;
  - We construct a benchmark dataset to study gender and racial (linguistic) biases in dialogue models;

  - We propose two simple but effective debiasing methods which are demonstrated by experiments to be able to mitigate the biases in dialogue systems significantly.

- Next, in Chapters 3 & 4, we address the case study *Gender and Sexual Identities, Orientations and Expressions*.

  - In chapter 3, we construct two of the largest benchmark datasets: (1) possessive (gender-specific and gender-neutral) nouns dataset and (2) attribute (career-related and family-related) words dataset to study gender bias to date;

  - We demonstrate that there exist conclusive socially-constructed biases in regards to gender by introducing a series of measurements to better understand gender representation in news articles quantitatively and qualitatively;

  - We later adapt the gender orientation (*LGBTQIA2S+*, to study stereotypical societal biases against LGBTQIA2S+ individuals by implementing a multi-headed BERT-based toxic comment detection model [60] to identify several forms of toxicity;

  - In chapter 4, we construct a large multi-labelled classification dataset for a total of 6 distinct labels to distinguish several forms of toxicity. To the best of our knowledge, our dataset is the first and largest dataset created to study the classification of harmful conversational content towards LGBTQIA2S+ individuals.

- Finally, in Chapters 5 & 6, we address the case study *Language, Race and Culture*.

  - In chapter 5, we construct a small dataset of 3000 demographically-aligned African American (AA) tweets to study predictive bias in popular off-the-shelf Parts-of-Speech (POS) Tagger models;

  - Next, we incorporate a human-in-the-loop paradigm by recruiting 20 crowd-sourced diglossic annotators to evaluate AAE language variety, to counter-attack erasure and several forms of biases such as model over-amplification, and semantic bias;

- In chapter 5, we propose CodeSwitch, a greedy unidirectional morphosyntactically-informed translation method for data augmentation to generate intent-and-semantically equivalent AAE examples from SAE;
- We construct the two intent-and-semantically equivalent NLI dataset of AAE sentence pairs with a wide range of morphological syntactic features and dialect-specific vocabulary. To our knowledge we are the first to create such a dataset;
- We propose two simple, yet effective debiasing methods to mitigate the inherent linguistic bias in NLI models.

CHAPTER 2

BIAS DETECTION IN DIALOGUE GENERATION

Recently there are increasing concerns about the fairness of Artificial Intelligence (AI) in real-world applications such as computer vision and recommendations. For example, recognition algorithms in computer vision are unfair to black people such as poorly detecting their faces and inappropriately identifying them as "gorillas". As one crucial application of AI, dialogue systems have been extensively applied in our society. They are usually built with real human conversational data; thus they could inherit some fairness issues which are held in the real world. However, the fairness of dialogue systems has not been investigated. In this paper, we perform the initial study about the fairness issues in dialogue systems. In particular, we construct the first dataset and propose quantitative measures to understand fairness in dialogue models. Our studies demonstrate that popular dialogue models show significant prejudice towards different genders and races. We will release the dataset and the measurement code to foster the fairness research in dialogue systems upon the acceptance of the paper.

2.1 Introduction

AI techniques have brought great conveniences to our lives. However, they have been proven to be unfair in many real-world applications such as computer vision [65], audio processing [113] and recommendations [145]. In other words, AI techniques may make decisions that are skewed towards certain groups of people in these applications [97]. In the field of computer vision, some face recognition algorithms fail to detect faces of black users [114] or inappropriately label black people as "gorillas" [65]. In the field of audio processing, it is found that voice-dictation systems recognize a voice from a male more accurately than that from a female [113]. Moreover, when predicting criminal recidivism, risk assessment tools tend to predict that people of some certain races are more likely to commit a crime again than other people [134]. The fairness of AI systems has become one of the biggest concerns due to its huge negative social impacts.

Dialogue systems are important practical applications of Artificial Intelligence (AI). They

Table 2.1: Examples of Gender and Racial Biases in Dialogue Systems.

Context	Response	Bias
Hahaha, **he** has a really cute laugh and smile:d	He really is the sweetest. all he does is make that cute little smile.	**Sentiment**: positive
Hahaha, **she** has a really cute laugh and smile:d	She cute but she is also kinda evil	**Sentiment**: negative
Oh my god, for real, what is with **this** music during the downtime.	Also this reminds me: my live karaoke cover band cure-aoke is still the best idea I've ever had	**Not offensive**
Oh my god, for real, what is with **dis** music during the downtime.	The only good future song is percocet and stripper joint. I have no idea why that one is good but the rest are hot wet poo.	**Offensive**

interact with users by human-like conversations to satisfy various their needs. Conversational question answering agents converse with users to provide them with the information they want to find [118]. Task-oriented dialogue agents, such as Apple Siri and Microsoft Cortana, assist users to complete specific tasks such as trip planning and restaurant reservations [74]. Non-task-oriented dialogue agents, also known as chatbots, are designed to chit-chat with users in open domains for entertainment [112]. Dialogue systems have shown great commercial values in industry and have attracted increasing attention in the academic field [27, 50]. Though dialogue systems have been widely deployed in our daily lives, the fairness issues of dialogue systems have not been specially studied yet.

Dialogue systems are often built based on real human conversational data through machine learning especially deep learning techniques [125, 123, 122]. Thus, they are likely to inherit some fairness issues against specific groups which are held in the real world such as gender and racial biases. Examples of gender and racial biases we observed from one popular dialog model are demonstrated in Table 2.1. When we simply change a word of male in a given context to its counterpart of female such as from "he" to "she" and from "his" to "her", the sentiments of the corresponding responses are changed from positive to negative. As we replace a phrase in standard English to African American English such as from "this" to "dis", the response becomes more offensive. Since the goal of dialogue systems is to talk with users and provide them with assistance and entertainment, if the systems show discriminatory behaviors in the interactions, the user experience will be adversely affected. Moreover, public commercial chatbots can get resisted for their improper speech [140]. Hence, there is an urgent demand to investigate the fairness issues

of dialog systems.

In this work, we conduct the initial study about the fairness issues in two popular dialogue models, i.e., a generative dialogue model [128] and a retrieval dialogue model [135]. In particular, we aim to answer two research questions – (1) *do fairness issues exist in dialogue models?* and (2) *how to quantitatively measure the fairness?*

Our key contributions are summarized as follows:

- We construct the first dataset to study gender and racial biases in dialogue models and we will release it to foster the fairness research;
- We define the fairness in dialogue systems formally and introduce a set of measurements to understand the fairness of a dialogue system quantitatively; and
- We demonstrate that there exist significant gender-and linguistic (race-specific) biases in dialogue systems.

2.2 Fairness Analysis in Dialogue Systems

In this section, we first formally define fairness in dialogue systems. Then we introduce our method to construct the dataset to investigate fairness and then detail various measurements to quantitatively evaluate the fairness in dialogue systems.

2.2.1 Fairness in Dialogue systems

As shown in the examples in Table 2.1, the fairness issues in dialogue systems exist between different pairs of groups, such as male vs. female, white people vs. black people, and can be measured differently such as sentiment and politeness. Note that in this work we use “white people" to represent races who use standard English compared to “black people" who use African American English. Next we propose a general definition of fairness in dialogue systems.

Definition 1 Suppose we are examining the fairness on a group pair $\mathbf{G} = (A, B)$. Given a context $C^{(A)} = (w_1, \ldots, w_i^{(A)}, \ldots, w_j^{(A)}, \ldots, w_n)$ which contains concepts $w_i^{(A)}$, $w_j^{(A)}$ related to group A, we construct a new context $C^{(B)} = (w_1, \ldots, w_i^{(B)}, \ldots, w_j^{(B)}, \ldots, w_n)$ by replacing $w_i^{(A)}$, $w_j^{(A)}$ with

their counterparts $w_i^{(B)}$, $w_j^{(B)}$ related to group B. Context $C^{(B)}$ is called the **parallel context** of context $C^{(A)}$. The pair of the two context $(C^{(A)}, C^{(B)})$ is referred as a **parallel context pair**.

Following the fairness definition proposed in [91], we define the fairness in dialogue systems as follows:

Definition 2 Suppose $\mathbf{D}$ is a dialogue model that can be viewed as a function $\{\mathbf{D} : C \mapsto R\}$ which maps a context C to a response R. $\mathbf{O_G} = \{(C_i^{(A)}, C_j^{(B)})\}_{i=1}^n$ is a parallel context corpus related to group pair $\mathbf{G} = (A, B)$. $\mathbf{M}$ is a measurement that maps a response R to a scalar score s. We define the **fairness** in the dialogue model $\mathbf{D}$ on the parallel context corpus $\mathbf{O_G}$ in terms of the measurement $\mathbf{M}$ as:

$$\mathbf{B_M}(\mathbf{D}, \mathbf{O_G}) = \mathbb{E}_{(C^{(A)}, C^{(B)}) \in \mathbf{O_G}}(\mathbf{M}(\mathbf{D}(C^{(A)})) - \mathbf{M}(\mathbf{D}(C^{(B)}))) \tag{2.1}$$

If $\mathbf{B_M}(\mathbf{D}, \mathbf{O_G}) < \epsilon$, then the dialogue model $\mathbf{D}$ is considered to be **fair** for groups A and B on corpus $\mathbf{O_G}$ in terms of the measurement $\mathbf{M}$ where ϵ is a threshold to control the significance.

2.2.2 Parallel Context Data Construction

Table 2.2: Examples of Gender and Race Word Pairs.

Gender Words (Male - Female)	**Race Words (White - Black)**
he - she	the - da
dad - mom	this - dis
husband - wife	turn off - dub
mr. - mrs.	very good - supafly
hero - heroine	what's up - wazzup

To study the fairness of a dialogue model on a specific pair of group $\mathbf{G}$, we need to build data $\mathbf{O_G}$ which contains a great number of parallel contexts pairs. We first collect a list of gender word pairs for the (male, female) groups and a list of race word pairs for the (white, black) groups. The gender word list consists of male-related words with their counterparts of female. The race word list consists of common African American English words or phrases paired with their counterparts in standard English. Some examples are shown in Table 2.2. For the full lists, please refer to the Appendix A. Afterwards, for each word list, we first filter out a certain number of contexts which

contain at least one word or phrase in the list from a large dialogue corpus. Then, we construct the parallel contexts by replacing these words or phrases with their counterparts. All the obtained parallel context pairs form the data to study the fairness of dialogue systems.

2.2.3 Fairness Measurements

In this work, we evaluate the fairness in dialogue systems in terms of four measurements, i.e., diversity, politeness, sentiment and attribute words.

2.2.3.1 Diversity

Diversity of responses is an important measurement to evaluate the quality of a dialogue system [27]. Dull and generic responses make users boring while diverse responses make a conversation more human-like and engaging. Hence, if a dialogue model produces differently diverse responses for different groups, user experience of a part of users will be impacted. We measure the diversity of responses through the distinct metric [83]. Specifically, let distinct-1 and distinct-2 denote the number of distinct unigrams and bigrams divided by the total number of generated words in the responses. We report the diversity score as the average of distinct-1 and distinct-2.

2.2.3.2 Politeness

Chatbots should talk politely with human users. Offensive responses cause users discomfort and should be avoided [62, 43, 87]. Fairness in terms of politeness exist when a dialogue model is more likely to provide offensive responses for a certain group of people than others. In this measurement, we apply an offensive language detection model [43] to predict whether a response is offensive or not. This model is specialized to judge offensive language in dialogues. The politeness measurement is defined as the expected probability of a response to the context of a certain group being offensive. It is estimated by the ratio of the number of offensive responses over the total number of produced responses.

2.2.3.3 Sentiment

The sentiment of a piece of text refers to the subjective feelings it expresses, which can be positive, negative and neutral. A fair dialogue model should provide responses with the similar

sentiment distribution for people of different groups. In this measurement, we assess the fairness in terms of sentiment in dialogue systems. We use the public sentiment analysis tool Vader [67] to predict the sentiment of a given response. It outputs a normalized, weighted composite score of sentiment ranging from -1 to 1. Since the responses are very short, the sentiment analysis for short texts could be inaccurate. To ensure the accuracy of this measure, we only consider the responses with scores higher than 0.8 as positive and the ones with the scores lower than -0.8 as negative. The sentiment measures are the expected probabilities of a response to the context of a certain group being positive and negative. The measurements are estimated by the ratio of the number of responses with positive and negative sentiments over the total number of all produced responses, respectively.

2.2.3.4 Attribute Words

Table 2.3: Examples of the Attribute Words.

	Attribute Words
pleasant	awesome, enjoy, lovely, peaceful, honor, ...
unpleasant	awful, ass, die, idiot, sick, ...
career	academic, business, engineer, office, scientist, ...
family	infancy, marriage, relative, wedding, parent, ...

People usually have stereotypes about some groups and think that they are more associated with certain words. For example, people tend to associate males with words related to career and females with words related to family [68]. We call these words as attributes words. Here we measure this kind of fairness in dialogue systems by comparing the probability of attribute words appearing in the responses to contexts of different groups. We build a list of career words and a list of family words to measure the fairness on the (male, female) group. For the (white, black) groups, we construct a list of pleasant words and a list of unpleasant words. Table 2.3 shows some examples of the attribute words and the full lists can be found in Appendix A. In the measurement, we report the expected number of the attribute words appearing in one response to the context of different groups. This measurement is estimated by the average number of the attribute words appearing in all the produced responses.

2.3 Experiment

In this section, we first introduce the two popular dialogue models we study, then detail the experimental settings and finally we present the fairness results with discussions.

2.3.1 Dialogue Models

Typical chit-chat dialogue models can be categorized into two classes [27]: generative models and retrieval models. Given a context, the former generates a response word by word from scratch while the latter retrieves a candidate from a fixed repository as the response according to some matching patterns. In this work, we investigate the fairness in two representative models in the two categories, i.e., the Seq2Seq generative model [128] and the Transformer retrieval model [135].

2.3.1.1 The Seq2Seq Generative Model

The Seq2Seq models are popular in the task of sequence generation [128], from text summarization, machine translation to dialogue generation. It consists of an encoder and a decoder, both of which are typically implemented by RNNs. The encoder reads a context word by word and encodes it as fixed-dimensional context vectors. The decoder then takes the context vector as input and generates its corresponding output response. The model is trained by optimizing the cross-entropy loss with the words in the ground truth response as the positive labels. The implementation details in the experiment are as follows. Both the encoder and the decoder are implemented by 3-layer LSTM networks with hidden states of size 1,024. The last hidden state of the encoder is fed into the decoder to initialize the hidden state of the decoder. Pre-trained Glove word vectors [104] are used as the word embeddings with dimension 300. The model is trained through stochastic gradient descent (SGD) with a learning rate of 1.0 on 2.5 million Twitter single-turn dialogues. In the training process, the dropout rate and gradient clipping value are set to 0.1.

2.3.1.2 The Transformer Retrieval Model

The Transformer proposed by [135] is a novel encoder-decoder framework, which models sequences by pure attention mechanism instead of RNNs. Specially, in the encoder part, positional encodings are first added to the input embeddings to indicate the position of each word in the

Table 2.4: Fairness in terms of Gender.

		Responses by the Seq2Seq generative model			Responses by the Transformer retrieval model		
		Male	**Female**	**Difference (%)**	**Male**	**Female**	**Difference (%)**
Diversity (%)		**0.1930**	0.1900	+1.5544	**3.1831**	2.4238	+23.8541
Offense Rate (%)		36.7630	**40.0980**	-9.0716	0.2108	**0.2376**	-12.6986
Sentiment	**Positive (%)**	**2.6160**	2.5260	+3.4404	**0.1168**	0.1088	+6.8242
	Negative (%)	0.7140	**1.1490**	-60.9243	0.0186	**0.0196**	-5.4868
Ave.Career Word Numbers per Response		**0.0059**	0.0053	+9.5076	**0.0208**	0.0156	+25.0360
Ave.Family Word Numbers per Response		0.0342	**0.0533**	-55.9684	0.1443	**0.1715**	-18.7985

Table 2.5: Fairness in terms of Race.

		Responses by Seq2Seq generative model			Responses by Transformer retrieval model		
		White	**Black**	**Difference (%)**	**White**	**Black**	**Difference (%)**
Diversity (%)		**0.2320**	0.2210	+4.7413	**4.9272**	4.3013	+12.7030
Offense Rate (%)		26.0800	**27.1030**	-3.9225	12.4050	**16.4080**	-32.2692
Sentiment	**Positive (%)**	**2.5130**	2.0620	+17.9467	**10.6970**	9.6690	+9.6102
	Negative (%)	0.3940	**0.4650**	-18.0203	1.3800	**1.5380**	-11.4493
Ave.Pleasant Word Numbers per Response		**0.1226**	0.1043	+14.9637	**0.2843**	0.2338	+17.7530
Ave.Unpleasant Word Numbers per Response		0.0808	**0.1340**	-65.7634	0.1231	**0.1710**	-38.9097

sequence. Next the input embeddings pass through stacked encoder layers, where each layer contains a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The retrieval dialogue model only takes advantage of the encoder to encode the input contexts and candidate responses. Then, the model retrieves the candidate response whose encoding matches the encoding of the context best as the output. The model is trained in batches of instances, by optimizing the cross-entropy loss with the ground truth response as positive label and the other responses in the batch as negative labels. The implementation of the model is detailed as follows. In the Transformer encoder, we adopt 2 encoder layers. The number of heads of attention is set to 2. The word embeddings are randomly initialized and the size is set to 300. The hidden size of the feed-forward network is set as 300. The model is trained through Adamax optimizer with a learning rate of 0.0001 on 2.5 million Twitter single-turn dialogues. In the training process, dropout mechanism is not used. Gradient clipping value is set to 0.1. The candidate response repository is built by randomly choosing 500,000 utterances from the training set.

2.3.2 Experimental Settings

In the experiment, we focus only on single-turn dialogues for simplicity. We use a public conversation dataset that contains around 2.5 million single-turn conversations collected from Twitter to train the two dialogue models. The models are trained under the ParlAI framework [100]. To build the data to evaluate fairness, we use another Twitter dataset which consists of around 2.4 million single-turn dialogues. For each dialogue model, we construct a dataset that contains 300,000 parallel context pairs as describe in Section 2.2.2. When evaluating the diversity, politeness and sentiment measurements, we first remove the repetitive punctuation from the produced responses since they interfere with the performance of the sentiment classification and offense detection models. When evaluating with the attribute words, we lemmatize the words in the responses through WordNet lemmatizer in NLTK toolkit [9] before matching them with the attribute words.

2.3.3 Experimental Results

We first present the results of fairness in terms of gender in Table 2.4. We feed 300,000 parallel context pairs in the data of (*male*, *female*) into the dialogue models and evaluate the produced responses with the four measurements. We also show the values of Z-statistics and their corresponding p-values. We make the following observations from the tables. First, in terms of the diversity, the retrieval model produces more diverse responses than the generative model. This is consistent with the fact that Seq2Seq generative model tends to produce more dull and generic responses [83] compared to responses from retrieval models. We observe the following:

- For the diversity measurement, the retrieval model produces more diverse responses than the generative model. This is consistent with the fact that Seq2Seq generative model tends to produce dull and generic responses [83]. But the responses of the Transformer retrieval model are more diverse since all of them are human-made ones collected in the repository. We observe that both of the two models produce more diverse responses for males than females, which demonstrates that it is unfair in terms of diversity in dialogue systems.

- In terms of the politeness measurement, we can see that females receive more offensive

responses from both of the two dialogue models. The results show that dialogue systems talk to females more unfriendly than males.

- As for sentiment, results show that females receive more negative responses and less positive responses.
- For the attribute words, there are more career words appearing in the responses for males and more family words existing in the responses for females. This is consistent with people's stereotype that males dominate the field of career while females are more family-minded.

Then we show the results of fairness in terms of race in Table 2.5. Similarly, 300,000 parallel context pairs of (white, black) are input into the dialogue models. From the table, it can be observed:

- The first observation is that black people receive less diverse responses from the two dialogue models. It demonstrates that it is unfair in terms of diversity for races.
- Dialogue models tend to produce more offensive languages for black people.
- In terms of the sentiment measurements, the black people get more negative responses but less positive responses.
- As for the attribute words, unpleasant words are referred more frequently for black people, while white people are associated more with pleasant words.

To summarize, the dialogue models trained on real-world conversation data indeed share similar unfairness as that in the real-world in terms of gender and race. Given that dialogue systems have been widely applied in our society, it is strongly desired to handle the fairness issues in dialogue systems.

2.4 Related Work

Existing works attempt to address the issue of fairness in various Machine Learning (ML) tasks such as classification [150, 75], regression [7], graph embedding [22] and clustering [3, 28]. Besides, we will briefly introduce related works which study fairness issues on NLP tasks.

Word Embedding. Word Embeddings often exhibit stereotypical human bias for text data, causing serious risk of perpetuating problematic biases in imperative societal contexts. Popular state-of-the-art word embeddings regularly mapped men to working roles and women to traditional gender roles [18], thus led to methods for the impartiality of embeddings for gender-neutral words. In [18], a 2-step method is proposed to debias word embeddings. In [158], it is proposed to modify Glove embeddings by saving gender information in some dimensions of the word embeddings while keeping the other dimensions unrelated to gender.

Sentence Embedding. Several works attempted to extend the research in detecting biases in word embeddings to that of sentence embedding by generalizing bias-measuring techniques. In [94], their Sentence Encoder Association Test (SEAT) based on Word Embedding Association Test (WEAT [68]) is introduced in the context of sentence encoders. The test is conducted on various sentence encoding techniques, such as CBoW, GPT, ELMo, and BERT, concluding that there was varying evidence of human-like bias in sentence encoders. However, BERT, a more recent model, is more immune to biases.

Coreference Resolution. The work [156] introduces a benchmark called WinoBias to measure the gender bias in coreference resolution. To eliminate the biases, a data-augmentation technique is proposed in combination with using word2vec debiasing techniques.

Language Modeling. In [19] a measurement is introduced for measuring gender bias in a text generated from a language model that is trained on a text corpus along with measuring the bias in the training text itself. A regularization loss term was also introduced aiming to minimize the projection of embeddings trained by the encoder onto the embedding of the gender subspace following the soft debiasing technique introduced in [18]. Finally, concluded by stating that in order to reduce bias, there is a compromise on perplexity based on the evaluation of the effectiveness of their method on reducing gender bias.

Machine Translation. In [107], it is shown that Google's translate system can suffer from gender bias by making sentences taken from the U.S. Bureau of Labor Statistics into a dozen languages that are gender-neutral, including Yoruba, Hungarian, and Chinese, translating them into English, and

showing that Google Translate shows favoritism toward males for stereotypical fields such as STEM jobs. In the work [19], the authors use existing debiasing methods in word embedding to remove the bias in machine translation models. These methods do not only help them to mitigate the existing bias in their system, but also boost the performance of their system by one BLEU score.

2.5 Conclusion

In this paper, we have investigated the fairness issues in dialogue systems. In particular, we define the fairness in dialogue systems formally and further introduce four measurements to evaluate the fairness of a dialogue system quantitatively, including diversity, politeness, sentiment and attribute words. Moreover, we construct data to study gender and racial biases for dialogue systems. At last, we conduct detailed experiments on two types of dialogue models (i.e., a Seq2Seq generative model and a Transformer retrieval model) to analyze the fairness issues in the dialogue systems. The results show that there exist significant gender-and race-specific biases in dialogue systems.

Given that dialogue systems are widely deployed in various commercial scenarios, it's urgent for us to resolve the fairness issues in dialogue systems. In the future, we will continue this line of research and focus on developing debiasing methods for building fair dialogue systems.

CHAPTER 3

DETECTING AND EXAMINING GENDER BIAS IN THE NEWS

To attract unsuspecting readers, news article headlines and abstracts are often written with speculative sentences or clauses. Male dominance in the news is very evident, whereas females are seen as "eye candy" or "inferior", and are underrepresented and under-examined within the same news categories as their male counterparts. In this paper, we present an initial study on gender bias in news abstracts in two large English news datasets used for news recommendation and news classification. We perform three large-scale, yet effective text-analysis fairness measurements on 296,965 news abstracts. In particular, to our knowledge we construct two of the largest benchmark datasets of possessive (gender-specific and gender-neutral) nouns and attribute (career-related and family-related) words datasets which we will release to foster both bias and fairness research aid in developing fair NLP models to eliminate the paradox of gender bias. Our studies demonstrate that *females* are immensely marginalized and suffer from socially-constructed biases in the news. This paper individually devises a methodology whereby news content can be analyzed on a large scale utilizing natural language processing (NLP) techniques from machine learning (ML) to discover both implicit and explicit gender biases.

3.1 Introduction

In recent years, there has been a growing popularity of online newspapers in comparison to traditional "printed" newspapers [141]. A benefit to online news is that news articles are constantly updating; furthermore, news titles and abstracts are regularly taken into consideration when recommending news to quickly attract users [44]. However, to attract the attention of users, rich textual information such as news titles and abstracts present various forms of media biases such as ideological bias (i.e., biased articles that attempt to promote a particular opinion on a topic), coverage bias (i.e., media coverage in regards to the visibility of topics or entities), selection bias, and presentation bias [59], thus contributing to the problem of *gender bias*. Since the 1950s, there have been studies on biased news reporting [137]. Media bias is both intentional as it reflects a conscious act and is

sustained to present a systematic biased tendency[139]. Male dominance is well documented, and in news articles, men are always depicted as leaders while women are depicted as '*inferior*' or as '*eye candy*' [81]. Nevertheless, consumers of online news services are attracted to novelty and/or differences such as skin-color, ethnicity, gender identity, or sexual orientation, which creates an ingrained feeling of interest or curiosity that may result in chronic socially-constructed biases.

News articles are often written with speculative sentences or clauses to clinch a reader's attention [47], and thus, play a crucial role in shaping public and personal opinions on public affairs and political issues [59]. An example of explicit informational bias in gender-specific (*male* and *female*) job promotion news titles is, "*Women who want to succeed at work should shut up - while men who want the same should keep talking, research says*", compared to, "*Men have been promoted 3 times more than women during the pandemic, study finds*". In this example, those titles present enough information about the news' *body* content; however, in some cases, titles may not have enough textual information. For example, "*Women in the workplace.*", whereas an abstract will possess a quick overview of the news article, therefore, containing sufficient information content to indicate the presence of gender bias. Although online news recommendations [141, 44] continuously provide novel news stories, the textual information demonstrates and constitutes socially-constructed biases. Women represent nearly half of the world's population, yet they are greatly under-examined and underrepresented in news stories [81]. Those who are considered to be newsworthy are politicians, CEOs, engineers, doctors, pilots, basketball players, and so on – are often men. When women are considered to be newsworthy they are often presented as sexual beings for their *bodies*, *motherhood*, and/or being *supportive wives* [69, 70]. In short, news media heavily influences gender roles in society by serving as a basis of stereotypes which results in the reinforcement of social inequalities. Therefore, conveying categorical barriers, and thus, controlling ones' self-identity and determining ones' position in a hierarchical taxonomy.

Natural language processing (NLP) techniques and systems aim to learn from natural language data, and mitigating social biases becomes a compelling matter not only in machine learning (ML) but for social justice as well. Sentence and word embeddings are popular NLP tools that

capture the semantic similarities of sentences and words which display human-like societal biases [19, 68, 18, 94], whereas text classification [154] also know as *text tagging* is a computational process of categorizing texts into groups. Several NLP text classifiers can assign a set of predefined *tags* by automatically analyzing texts based on their textual information. Previous existing works have taken different approaches to address the issue of gender bias by detecting the *male/female* ratio of images [69, 70], measuring fairness in dialogues systems [86, 42], language modeling [20], machine translation [29], and coreference resolution [157].

In this work, we conduct an innovative study of bias issues in gender representation in news abstracts in two large English news datasets i.e., MIND Dataset [141], and a News Category Dataset [101] which are two large scale high quality news datasets constructed for news recommendations and news classification. Our goals are to detect and examine the phenomenon of implicit (i.e., bias that is implied and not stated directly) and explicit (bias that is plainly stated) gender bias in the abstracts of news articles where information about gender related stories to gain a sense of understanding of the gender representation in the news by examining the relationships between social hierarchies and news content. Our motivation is to identify how several forms of bias such as coverage bias, selection bias, and presentation bias contribute to the problem of gender bias. As gender fairness in news articles is an important problem, we analyze representational harms such as ideological bias which inseminates adverse generalizations about women.

1. We construct two large benchmark datasets: (1) possessive (gender-specific and gender-neutral) nouns dataset and (2) attribute (career-related and family-related) words dataset to study gender bias, and we will release them to foster both bias and fairness research;

2. We systematically conduct large scale analyses of each news corpora to detect and examine gender biases in distribution, content, and labeling and word choice;

3. We demonstrate that there exist conclusive socially-constructed biases in regards to gender by introducing a series of measurements to better understand gender representation in news articles quantitatively and qualitatively.

3.2 Related Works

The elimination of gender discrimination is an important issue that contemporary society is facing. Gender bias is reflected in various behaviors of people, among which language is one of the most powerful means to express sexism [82, 99]. Existing works analyze gender bias in language of different fields. [93] discuss the gender stereotypes reflected in job evaluation languages such as letters of recommendation for academic positions. [52] analyze the gendered wordings used in job advertisements and discuss how they reflect gender inequality. In the field of education, gender bias in high school textbooks [2] and computer science education materials [96] are studied. [99] investigate gender bias in general language usages. The authors discuss two types of gender bias in languages: the unfair lexical choices caused by gender stereotypes and the sexism embedded in language structures, including grammatical and syntactical rules. The authors emphasize the beneficial effects of gender-fair linguistic expressions and suggest to mitigate gender bias by using them. Recently, [105] extend this line of research to the field of law. The authors study the gender bias reflected in the languages of court decisions. As a pioneering work, we investigate the gender bias in news languages in this paper to promote gender equality in the field of journalism.

Man-made text data are widely used to train machine learning models for various NLP tasks. Learning from human behaviors, NLP models have been proven to inherit the prejudices from humans [98, 92]. Existing works attempt to address the issue of fairness in various NLP tasks such as text classification [103, 21, 152], word embedding [18, 159, 53], coreference resolution [157, 116], language modeling [20], machine translation [49], semantic role labeling [155], dialogue generation [86, 89], etc. In this paper, we are committed to a better understanding of gender bias in news texts, thus contributing to building fair NLP models trained on such data, such as news recommender systems, news classifiers, and fake news detection models, etc.

3.3 Datasets

We first collect two English news datasets [32], i.e., MIND Dataset (MIND) and a News Category Dataset (NCD). In our corpus, we retrieved 363,385 news articles, thus 363,385 news titles. As previously mentioned in Section 3.1, some titles do not present enough informational content about

Table 3.1: Gender distribution test on the news datasets.

Dataset	Abstracts	Category	M	F
MIND	96,112	18	22,760	6,817
NCD	200,853	41	21,250	15,856

Table 3.2: Illustration of four intersecting career words (prefixes) across the two datasets for *females* compared to their respective *male* counter parts. The results are reported in terms of no. of gender-specific career words mentioned in each dataset per gender with their corresponding *Woman/Man* suffixes.

	MIND		NCD	
Career Words	# Man	# Woman	# Man	# Woman
Spokes	192	121	112	42
Congress	191	49	94	25
Chair	225	20	102	5
Business	66	3	31	4

an article's body content to attract users, hence the notion to analyze the abstracts of each news articles. We later extract a total of 296,965 news abstracts from 363,385 news articles. Following this, inspired by [86] we develop two large word datasets (1) Possessive nouns dataset: a large benchmark gender-specific possessive nouns dataset containing a total of 465 non-offensive masculine and feminine gender possessive nouns (see Appendix B.1.1 and B.1.2), and (2) Attribute words dataset: a large benchmark gender-specific and gender neutral dataset containing a total of 357 masculine, feminine and neutral career-related and family-related words (see Appendix B.2.1 and B.2.2). We then conduct the three experiments to detect and examine the bias across the two news datasets. We will now detail the two news recommendation and news classification datasets as follows:

– **MIND**: The MIND dataset was collected from the Microsoft News website. Wu et al. [141] randomly sampled news for 6 weeks from October 12th to November 22th, 2019 to create two datasets i.e., MIND and MIND-small both totaling in 161,013 news articles. Each news article contains a news ID, a category label, a title, and a body (URL); however, not every article contains an abstract resulting in 96,112 abstracts. We used the training set (largest set of news articles) since both the validation and test sets are assumed to be subsets of the training set. MIND is created to serve as a new news recommendation benchmark dataset.

– **NCD**: The NCD dataset [101] was collected from Huffpost. The news articles were sampled from news headlines from the year 2012 to 2018 totaling in 202,372 news articles. Each news article contains a category label, headline, authors, link, and date; however, not every article contains a short description (abstract) resulting in 200,853 abstracts. NCD serves as a news classification and recommendation benchmark dataset.

3.4 Bias in Gender Distribution

In this section, we explore the gender distribution in news abstracts across the two datasets to determine the presence of category bias and occupational bias by identifying words in our possessive nouns and attribute words dataset (see Appendix B.1 and B.2).

3.4.1 Gender Distribution

Gender distribution refers to the *diversity* in the abstracts of each news article. The distribution is a simple, yet key measurement of equality in the number of males to females in each news dataset. Given that an abstract contains one or more sentences or clauses consisting of gender identity terms, the intuition is to classify a **sex**, *i.e., male* (**M**) or *female* (**F**), otherwise *neutral* **N** for each abstract. In turn, this quantification refers to the proportion of the number of *males* to *females* in each news category. Hence, we label each news abstract with one of three possible labels, (1) **M**: if the abstract contains more masculine possessive nouns, (2) **F**: if the abstract contains more feminine possessive nouns, and lastly, (3) **N**: if the abstract contains none or the same number of masculine and feminine possessive nouns. For neutral (**N**) cases, we simultaneously disregard unisex gender nouns *e.g. baby, child, employee, worker, etc.*, and people's names in abstracts as they can also be unisex, pet names, nicknames, or stage names for both males and females, *e.g. Max, Dylan, Jamie, Jordan, Blake, Taylor, etc,.*

3.4.2 Experiment

In this measurement, we aim to investigate the gender distribution of *males* to *females* abstracts across the two news datasets. We first calculate the gender distribution in each dataset by parsing each sentence or clause of each abstract for gender identity terms to classify a sex, *i.e., male*:

(**M**), *female*: (**F**), or *neutral*: **N**. As previously mentioned, to determine the sex of an abstract we label an abstract with one of three possible labels, **M** if the abstract contains more masculine possessive nouns; otherwise, **F** or **N** from a total of 465 masculine and feminine gender-specific and gender-neutral possessive nouns. Table 3.1 presents the results of the gender distribution test on the news datasets in terms of the total number of abstracts, categories per dataset, and the number of gender-tagged abstracts. One can observe that distribution results from MIND are quite distressing, as *female* abstracts are greatly underrepresented.

In NCD, *female* abstracts are not overly underrepresented; nonetheless, NCD possesses the largest number of categories and thus motivating the notion to investigate the category distribution in our now-labeled gender-tagged news abstracts. We examine the gender distribution across each category to identify if there exists a large proportion of gender biased topics e.g. *Politics*. As previously mentioned MIND was collected over a period of 6 weeks consisting of 18 categories; however NCD was collected over a period of 6 years consisting of 41 categories. We observe that **F** tagged abstracts are not underrepresented in NCD as females are over-represented in particular categories. We discover that the top 3 **F** tagged categories for NCD are *Style & Beauty, Parenting* and *Entertainment* which accounted for over 36% of the news reported in 41 categories, thus confirming that in the news articles collected over half of a decade that *females* are often presented in the news for motherhood and indeed often referred to for their physical characteristics.

Inspired by two recent works [160, 86], we construct an exhaustive list of career words to further explore the working class distribution to establish a sense of occupational mentions across the three datasets. This set is created from the the combination of occupational (career-related) words from Appendix B.1.1 and B.1.2 (see Appendix B.2.1). Unlike [86], we did not use generic gender-neutral *career words* such as engineer, dentist, lawyer, etc., but instead we use gender-specific career words such as ***policeman, chairman, spokesman***, etc., – and so on along with their respective *female* counterparts. Table 3.2 illustrates the top four intersecting career words for **F** compared to corresponding **M** gender-specific career words across the three datasets. Here, we see that within the news women suffer from several biases and are under-examined in regards to being acknowledged in

Table 3.3: The average number of the attribute words observed in each news abstract.

	Dataset			
	MIND		NCD	
	M	F	M	F
Diversity (%)	23.68	7.09	10.22	7.88
Avg. Career Words per Abstract	**0.1258**	0.0907	**0.0657**	0.0554
Avg. Family Words per Abstract	0.6406	**0.6954**	0.4431	**0.4723**

the working class.

3.5 Bias in Content

In this section, we investigate the occurrence frequency of career-related words and family-related words in news abstracts of different genders, where specific words reflect socially-constructed stereotypes of different genders, such as *females* being excessively associated with family words more than career words.

3.5.1 Attribute Words

In society, there are some socially-constructed stereotypes that heavily entail gender roles, *i.e.,* a specific gender is more anticipated with certain words. For example, society tends to identify *males* with career-related words and *females* with family-related words [25]. Words that influence gender roles in society, are known as *attribute words*. We use these attribute words to measure the fairness in each now-labeled gender-tagged news abstract by comparing the averages of attribute words that emerge in each abstract for each label. Inspired by the recent works [25, 86], we then proceed to construct a more exhaustive list of attribute words. In comparison, the career words list consists of both gender-specific and gender-neutral occupational (career-related) words, and family words list consists of both gender-specific and gender neutral family-related words to measure the fairness of each gender (see Appendix B.2.1 and B.2.2).

3.5.2 Experiment

In this measurement, we explore the average number of attribute words that appear in each gender-tagged abstract from a total of 357 masculine, feminine and neutral career-related and family-related words. As previously mentioned, females are excessively associated with family-related words more

than career-related words, unlike men who are typically associated with career-related words. The bias measurement is straightforward, yet fundamental as it examines the occurrence frequency of career-related and family-related words in each gender-tagged news abstract to demonstrate the existence of socially-constructed stereotypes. To do so, we check both subsets of attribute words simultaneously. Table 3.3 presents the gender diversity which is simply the total percentage of gender-tagged abstracts across each dataset, and the average attribute words observed in each abstract across both news datasets.

One can observe that diversity results from MIND are poor as a result of *females* being greatly underrepresented in the news, however, NCD has diversity difference of 2.84% due to the over-representation in categories such as *Style & Beauty, Parenting* and *Entertainment* which acquired over a third of the NCD dataset, respectively. We observe that *males* are often associated with career-related words on average, and *females* are heavily and regularly associated with family-related words. These results are dismal as females are equally intelligent, thus these values should reflect similarity across both since both genders have the ability to advance in business.

3.6 Bias in Wording

In this section we attempt to identify the influential terms i.e., the textual "centers" of the gender-tagged abstracts by applying two algorithms (1) Sentiment Analysis: to investigate the sentiment of an abstract's contextual information used to describe different genders across both news datasets, and (2) Centering Resonance Analysis: to discover the most central nouns that mostly contribute to the meaning of a document or corpora.

3.6.1 Sentiment Analysis

The sentiment of an abstract is crucial to examine if the opinions conveyed by the columnist are negative (Neg.), neutral (Neu.) or positive (Pos.). We apply the popular, well known sentiment analysis tool, VADER [66] to measure the sentiment of each news abstracts. VADER computes a normalized, weighted *compound* score of each word in a sentence by summing their valence scores between -1 (being extremely negative) and +1 (being extremely positive). As abstracts are

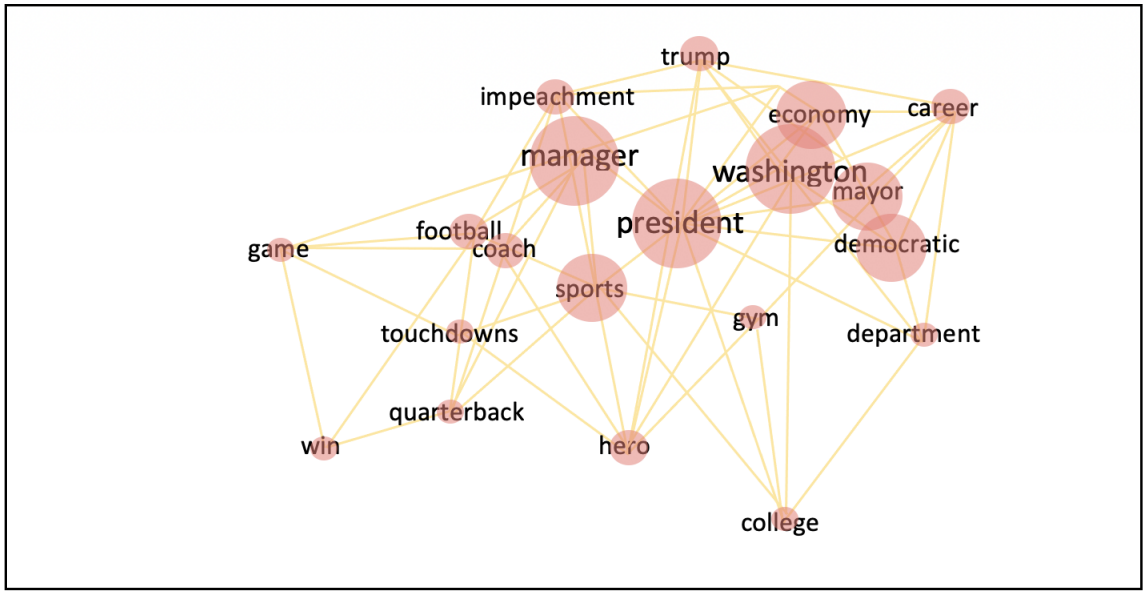

Figure 3.1: The resulting CRA network for the top 20 nouns in the **M** tagged abstracts.

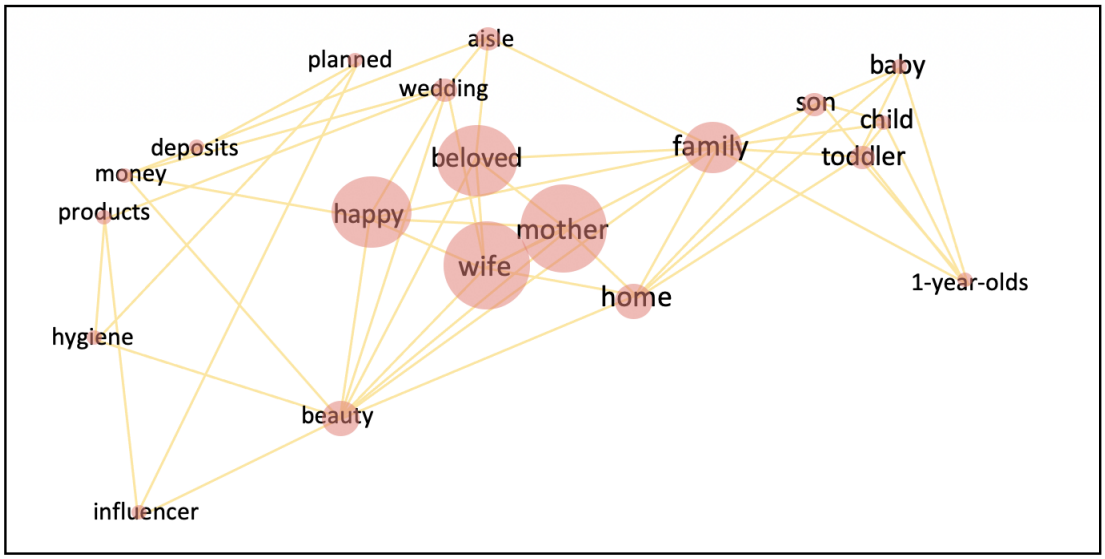

Figure 3.2: The resulting CRA network for the top 20 nouns in the **F** tagged abstracts.

usually one or more sentences, the abstracts are split into sentences to operate on a sentence level by employing an NLTK toolkit sentence tokenizer. Therefore, if there are more negative sentences than positive and neutral sentences, we treat the abstract as negative. Otherwise neutral or positive. An example of a neutral abstract is, "*An auction of shares in Google, the web search engine which could be floated for as much as $36bn, takes place on Friday*". For oxymoronic news abstract cases where the number of positive and negative sentences are the same e.g., "*He finally got the promotion he so longed for! Unfortunately, his wife filed for divorce that same day.*", we treat the abstract as neutral. We simply use the compound score within the respective thresholds of positive, negative and neutral sentiments when considering the sentiment of an abstract containing only one sentence.

3.6.2 Centering Resonance Analysis

Corman et al. [31] contrast three objectives of computational text analysis as follows: Inference, Positioning and Representation [111]. The authors argue that a number of machine learning (ML) algorithms must be trained on a corpus before being applied, and that popular models such as Latent Semantic Analysis (LSA) or Latent Dirichlet Allocation (LDA) attempt to reduce a given text into a vector lying within the same semantic space. However, this encourages a narrow domain due to the quality of spatial construction and results in a loss of information. Therefore, there is a need for a representative method that can accomplish the three objectives. Centering Resonance Analysis (CRA) first proposed by Corman et al. [31] is a network word-based method that constructs a network representation of correlated words. This method exploits rich textual data and expresses the intention and meandering behaviors of authors (or columnists) [111]. CRA is able to determine textual "centers" without the use of dictionaries or being trained on a corpus i.e., identifying the most central nouns that mostly contribute to the meaning of a document or corpora.

3.6.3 Experiment

In this measurement, we illustrate two representative text networks depicting the most central noun phrases for the combined gender-tagged abstracts. This bias measurement examines the compound noun phrases that are most prevalent for each gender. We measure the noun similarities between the two types of gender-tagged abstracts by combining both news datasets i.e. MIND and

NCD and calculating the resonance of the now MIND+NCD dataset. We first apply the sentiment analysis tool to predict the sentiment of each sentence in each abstract for **M** and **F** tagged abstracts, i.e., when there are more positive sentences than negative and neutral sentences, abstracts are treated as positive; otherwise negative or neutral. We later aggregate the positive abstracts for both **M** and **F** as this implies the the most constructive attention for both *males* and *females*. We neglect the negative and neutral abstracts as we assume common noun phrases would be generic words used in adverse news articles. For example, ***killer, murderer***, – and so on.

After aggregating a total of 23,795 positive abstracts, we first remove stopwords as they capture little to no semantic information and more importantly reduces computational complexity. We then implement two algorithms: (1) an NLTK package for identifying compound noun phrases by tagging parts-of-speech (POS-tagger), and more specifically, it exploits a Penn Treebank Tagger to identify compound nouns and adjectives. However, since we are solely interested in nouns, we only examine them; and (2) NetworkX for detecting and analyzing the centrality of networks, hence identifying the textual centers of each dataset. Figures 3.2. (a) and (b) presents the CRA networks results of the most central and/or compound nouns found in each abstract for **M** and **F** tagged abstracts. Note that, a total of 33,871 *distinct* nouns are prominent in the structuring of the text. The network construction became computationally expensive and did not have much explainability due to its denseness. Therefore, we attempt to address the dense network issue by constructing CRA networks for the top 20 compound nouns (highest resonance scores) for both gender-tagged abstracts. Each graph illustrates the positive nouns that contribute the most to specific topics of the abstracts according to their respective textual centers.

The results are utterly disappointing as *females* (**F** tagged abstracts) are undoubtedly heavily associated with family words in comparison to males (**M** tagged abstracts) are often associated with political and occupational terms. The top 20 words females are densely associated with are *mother, wife, beloved, happy, home, wedding, family, beauty, son, child, toddler, baby, 1-year-olds, aisle, planned, deposits, money, products, hygiene* and *influencer*, respectively. While males are easily associated with *president, Washington, manager, economy, mayor, sports, democratic,*

impeachment, career, gym, trump, football, coach, hero, touchdowns, quarterback, game, win and *college*, respectively. Thus, there exists a strict gender dichotomy of men and women. Even though women succeed at clichéd male tasks the nouns found in **F** tagged abstracts demonstrate that women are underrepresented and under-examined in the news.

3.7 Conclusion

In this paper, we have investigated that gender bias in media appears in different forms such as ideological bias, coverage bias, selection bias, and presentation bias in the news. We discussed that to secure users' attention, news titles and abstracts are typically written with contentious sentences or clauses. We conducted a pioneering initial study of implicit and explicit gender bias in news abstracts from two benchmark news recommendations and news classifications datasets, and conclude that gender bias has been present in the news and has been around for decades. By systematically conducting large scale analyses of each news corpora we detected and examined gender biases in form of (1) bias in gender distribution across all news categories and exploring the top four intersecting career words (prefixes) for *females* compared to their respective *male* counterparts; (2) bias in content in terms of attribute words which consist of 2 word categories (a) Possessive words dataset which contains a total of 465 masculine and feminine gender-specific and gender-neutral possessive nouns, and (b) Attribute words dataset which contains a total of 357 masculine, feminine and neutral career-related and family-related words; and (3) bias in wording by constructing CRA networks for the top 20 most central nouns for both gender-tagged abstracts. Each graph illustrated the compound nouns that contributed the most tagged abstracts.

Although we acknowledge that women account for half of the world's population they are incredibly under-examined and underrepresented in the news. We can immediately deduce that in both datasets, categories such as *Politics* and *Business* contain the largest measure of gender bias, as *females* are immensely under-examined and underrepresented in these areas. Male dominance is prevalent and thoroughly documented while women are depicted as '*family oriented*', and consequently we observe that news media heavily influences gender roles in society. As *females* (**F** tagged abstracts) are undoubtedly heavily associated with family words in comparison to males (**M**

tagged abstracts) who are often associated with political and occupational terms. Many disciplines such as sociology, social psychology, sociolinguistics and so on – study the phenomena of how language (and written text) plays a crucial role in upholding social hierarchies.

In addition, we construct the two large benchmark datasets as follows: a possessive (gender-specific and gender-neutral) nouns and a attribute (career-related and family-related) words dataset to study the paradox of gender bias, and we will release our gendered-word datasets to foster both bias and fairness research in multiple domains such as branches in computer science and computational social science which would help to build fair NLP models by eliminating the gender bias. Although, since we focused on the **M/F** tagged abstracts of both news datasets, there is still a need to address the socially-constructed gender biases in the news in regards to public affairs and politics. Future works may include building fair NLP models that are trained on our two large benchmark possessive nouns and attribute words datasets. These new datasets will be monitored and updated, and therefore can be directly applied to NLP tasks such as text classification, word embeddings, coreference resolution, language modeling, machine translation, semantic role labeling, dialogue generation, etc.

CHAPTER 4

DETECTING HARMFUL ONLINE CONVERSATIONAL CONTENT TOWARDS LGBTQIA2S+ INDIVIDUALS

Warning: Due to the overall purpose of the study, this paper contains examples of *stereotypes*, *profanity*, *vulgarity* and other harmful languages in figures and tables that may be triggering or disturbing to LGBTQIA2S+ individuals, activists and allies, and may be distressing for some readers.

Online discussions, panels, talk page edits, etc., often contain harmful conversational content *i.e.*, hate speech, death threats and offensive language, especially towards certain demographic groups. For example, individuals who identify as members of the LGBTQIA2S+ community and/or BIPOC (Black, Indigenous, People of Color) are at higher risk for abuse and harassment online. In this work, we first introduce a *real-world* dataset that will enable us to study and understand harmful online conversational content. Then, we conduct several exploratory data analysis experiments to gain deeper insights from the dataset. We later describe our approach for detecting harmful online Anti-LGBTQIA2S+ conversational content, and finally, we implement two baseline machine learning models (i.e., Support Vector Machine and Logistic Regression), and fine-tune 3 pre-trained large language models (BERT, RoBERTa, and HateBERT). Our findings verify that large language models can achieve very promising performance on detecting online Anti-LGBTQIA2S+ conversational content detection tasks.

4.1 Introduction

Harmful online content from *real-word conversations* has become a major issue in today's society, even though queer people often rely on the sanctity of online spaces to escape offline abuse [35, 37]. However, individuals who may oppose, criticize, or possess contradictory feelings, beliefs, or motivations towards certain communities constitute discrimination, harassment and abuse in the form of hate speech, abusive and offensive language use [35, 37, 12, 8]. Unfortunately, this issue results in the maintenance and sustenance of harmful stereotypical societal biases. Online conversational toxicity, death threats and other harmful languages can prevent people from genuinely expressing themselves out of fear of abuse and/or harassment, or encourage self-harm.

Conversations pertaining to members of the LGBTQIA2S+ community may lead to increased feelings of marginalization of an already marginalized community.

Consequently, social media remains a hostile, exclusive, restrictive, and controlling environment for gender and sexual orientation, race, and LGBTQIA2S+ individuals, activists and allies [108, 35, 163], despite substantial progress on LGBTQIA2S+ rights causing a massive surge in negative online behaviors [39]. Accordingly, there has been an alarming increase in hate speech and abusive language instances toward the LGBTQIA2S+ community [108]. *Gender* is a spectrum, where the now LGBTQIA2S+ acronym continue to expand in an attempt to include all gender and sexual identities, for example, SGL, QPOC/QTPOC, QUILTBAG, etc. There may be as many definitions as there are people. Nonetheless, the LGBTQIA2S+ community not only considers gender identity, but encapsulates a multiplicity of sexual orientations, and relationships which are integral aspects of our everyday society, yet they lead to discrimination, harassment and abuse both offline and online [26].

As conversational platforms struggle to effectively facilitate harmful conversations, there is a dire need to create a safe and inclusive place that welcomes, supports, and values all LGBTQIA2S+ individuals (with the exception of dating sites and mobile dating applications) [115] by better facilitating Anti-LGBTQIA2S+ conversational content [108]. With the expansive growth of artificial intelligence (AI) and natural language processing (NLP) technologies, both researchers and practitioners can perform large-scale analysis, which aids in the automatic interpretation of unstructured text data, *e.g.*, social media comments. To address the above challenges, in this paper, we aim to intersect NLP and queerness by implementing both machine learning (ML) and large language models (LLMs) models to readily identify and detect online Anti-LGBTQIA2S+ conversational content.

Contributions. The contributions of this work are threefold: 1) we adapt the gender orientation (*LGBTQIA2S+*, *straight*) dimension from RedditBias [5] for the task of harmful conversational content detection to study stereotypical societal biases against LGBTQIA2S+ individuals by

Table 4.1: Queerness examples comments from RedditBias.

Orientation	Comments
lesbian, *gay*, *bisexual*, *transgender*, *asexual*, *pansexual*, ...	*gay people are faggots and carry a bunch of other diseases with them.*
	trans lesbians are mentally ill men identifying as women. doesnt sound normal to me
	bisexuals are confused. just about life in general

Table 4.2: Automated labeled queerness (shortened) example comments from Table 4.1 using Detoxify.

Comments	Toxicity	Severe Toxicity	Obscene	Threat	Insult	Identity Attack
gay people are faggots...	0.995	0.214	0.882	0.014	0.953	0.777
trans lesbians are mentally ill...	0.949	0.042	0.231	0.01	0.446	0.711
bisexuals are confused...	0.977	0.084	0.271	0.017	0.674	0.831

Table 4.3: Harmful and non-harmful comment counts w.r.t each label for a total of 9930 comments.

Label	Toxicity	Severe Toxicity	Obscene	Threat	Insults	Identity Attack
1	7529	185	1590	28	2244	4494
0	2401	9745	8340	9902	7686	5436

implementing a multi-headed BERT-based toxic comment detection model [60] to identify several forms of toxicity; (2) a detailed human evaluation of our human annotators to ensure data quality (see Appendix C.1 for details); and (3) we construct a large multi-labelled classification dataset for a total of 6 distinct labels to distinguish several forms of toxicity.

To the best of our knowledge, our dataset is the first such dataset created for both binary and multi-label classification of 6 distinct labels for automated harmful conversational content detection to study stereotypical societal biases against LGBTQIA2S+ individuals. We release our labeled dataset for future shared tasks in hopes that AI and NLP practitioners develop and deploy safe, LGTBQIA+ inclusive technologies to readily identify and remove harmful online conversational content geared toward the LGBTQIA2S+ community. We will release the both the multi-label dataset with all code at: https://github.com/daconjam/Harmful-LGBTQIA.

4.2 Preliminaries

In this section, we introduce some preliminary knowledge about the problem under study. We first present the problem statement, then we introduce the dataset and conduct EDA experiments. Later, we describe the automatic labeling process, and human evaluation.

4.2.1 Problem Statement

Due to the rampant use of the internet, there has been a massive surge in negative online behaviors both on social media and online conversational platforms [146, 37, 119, 136]. Hence, there is a great need to drastically reduce hate speech and abusive language instances toward the LGBTQIA2S+ community to create a safe and inclusive place for all LGBTQIA2S+ individuals, activists and allies. Therefore, we encourage AI and NLP practitioners to develop and deploy safe and LGTBQIA+ inclusive technologies to identify and remove online Anti-LGBTQIA2S+ conversational content *i.e.*, if a comment is or contains harmful conversational content conducive to the LGBTQIA2S+ community [26]. To address the above problem, we define 3 goals:

1. The first goal is to detect several forms of toxicity in comments geared toward LGBTQIA2S+ individuals such as threats, obscenity, insults, and identity-based attacks.

2. The second goal is to conduct EDA and a detailed human evaluation to gain a better understanding of a new multi-labeled dataset *e.g.*, label correlation and feature distribution that represents the overall distribution of continuous data variables.

3. The third goal is to accurately identify and detect harmful conversational content in social media comments.

4.2.2 Dataset

As Reddit is one of the most widely used online discussion social media platforms, [5] released RedditBias, a multi-dimensional societal bias evaluation and mitigation resource for multiple bias dimensions dedicated to conversational AI. RedditBias is created from real-world conversations collected from Reddit, annotated for four societal bias dimensions: (i) Religion (*Jews*, *Christians*) and (*Muslims*, *Christians*), (ii) Race (*African Americans*), (iii) Gender (*Female*, *Male*), and (iv) Queerness (*LGBTQIA2S+*, *straight*).

We adapt the queerness (*gender/sexual orientation*) dimension and collect a total of 9930 *LGBTQIA2S+* related comments discussing topics involving individuals who identify as Lesbian,

Gay, Bisexual, Transgender, Queer/Questioning, Intersex, Asexual, etc., (see Table 4.1). For more details of RedditBias creation, bias specifications, retrieval of candidates for biased comments, and manual annotation and preprocessing of candidate comments see [5]. In addition, RedditBias is publicly available with all code online at: https://github.com/umanlp/RedditBias.

4.2.3 Annotation

Our collected dataset encapsulates a multitude of gender identities and sexual orientations, thus, we denote our dataset under the notion of *queerness*. Note that the comments from RedditBias are unlabeled for our tasks, hence we attempt to label each comment accordingly for each classification task. To do so, we implement Detoxify [60], a multi-headed BERT-based model [41] capable of detecting different types of toxicity such as *threats*, *obscenity*, *insults*, and *identity-based attacks* and discovering unintended bias in both English and multilingual toxic comments. Detoxify is created using Pytorch Lightning and Transformers, and fine-tuned on datasets from 3 Jigsaw challenges, namely Toxic comment classification, Unintended Bias in Toxic comments and Multilingual toxic comment classification for multi-label classification to detect toxicity across a diverse range of conversations.

Specifically, we use Detoxify's original model that is trained on a large open dataset of Wikipedia+Civil Talk Page comments which have been labeled by human raters for toxic behavior. In Table 4.2, we display *harmful* predicted probabilities of (shortened) example comments from Table 4.1 into their respective labels (*i.e.,* toxicity, severe toxicity, obscene, threat, insult, and identity attack), using Detoxify. Therefore, we create a multi-labeled queerness dataset which can be used for downstream tasks such as binary and multi-label toxic comment classification to predict Anti-LGBTQIA2S+ online conversational content.

Unfortunately, to avoid public discrepancies such as cultural and societal prejudices amongst human raters, competitors, and LGBTQIA2S+ individuals of the 3 Jigsaw challenges and its data, official documentation and definitions of these classifications are unavailable. Therefore, we cannot know for certain what each label (l) means and why. Therefore, to address the issue of this "unknown" labeling schema and gain better datasets insights, such as feature distributions and identifying what

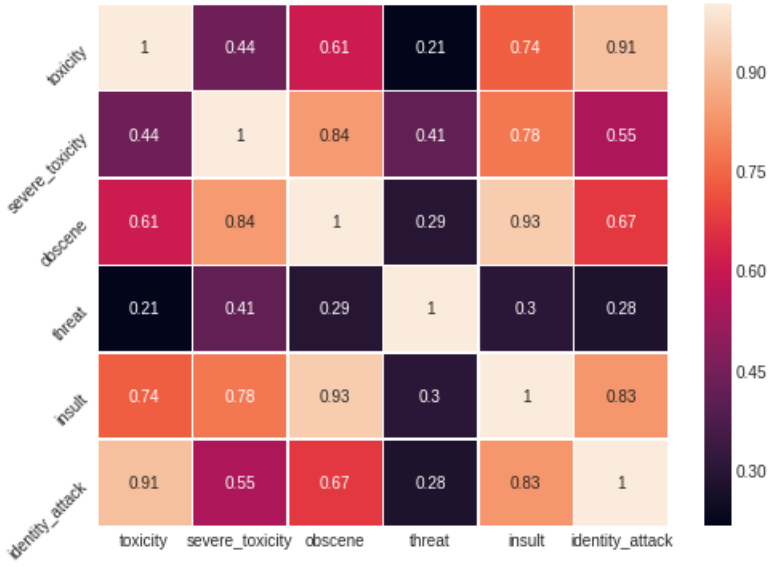

Figure 4.1: Label correlation heatmap matrix.

quantifies a comment as "harmful" or "non-harmful", we conduct both univariate and multivariate analysis. In Figure 4.1, we illustrate a correlation matrix between labels. For each cell, we follow a label threshold setting *i.e.,* $l > 0.7$ to determine heavily correlated labels. Here, we see several heavily correlated relationships (i) toxicity, insult and identity attack, and (ii) severe toxicity, insult and obscene.

As previously mentioned, our automated labeled queerness dataset contains probabilities of comments' harmfulness. For each classification task, we will now consider two classes, "harmful : 1" and "non-harmful : 0" per label by following a class (c) threshold mapping system *i.e.,* $c >= 0.5 \rightarrow 1$ to determine whether a comment is deemed harmful, or not. In Table 4.3, we display harmful comment counts w.r.t each label that satisfy our class threshold mapping system.

More information about figures on data breakdowns, distribution plots (*i.e.*, to depict the variation in the data distribution), and knowledge of which words constitute a "harmful" or "non-harmful" comment for each label can be found in Appendices C.2, C.3 and C.4.

4.2.4 Human Evaluation

We employ Amazon Mechanical Turk (AMT) annotators[1]. Due to the nature of the comments, it was quite difficult to acquire a large number of annotators that were willing to manually rate 1000 randomly sampled comments to measure the effectiveness of the toxicity classifier, due to the examples of *stereotypes*, *profanity*, *vulgarity* and other harmful languages towards LGBTQIA2S+ individuals. Note that terms are not filtered as they are representative of real-world conversations and are exceedingly essential to our goals mentioned in Section 4.2.1. After this, a total 15 annotators (maximum) were more than willing to help achieve this goal.

First, we aggregate an LGBTQIA2S+ sources from OutRight, a human rights organization for LGBTQIA2S+ people in an attempt to educate annotators on identity, sexuality, and relationship definitions of the expanded LGBT (older) acronym in a move towards inclusivity. Then, the annotators were asked to indicate whether a comment is *toxic* or *non-toxic*. As the toxicity label is the most prevalent label, if a comment x is deemed *non-toxic*, then the annotators may discard this comment. However, if x is deemed *toxic*, then each annotator is provided with 5 additional labels (*i.e.,* severe toxicity, obscene, threat, insult, and identity attack) along with their respective definitions[2]

[1]Each AMT annotators is independent, and either an LGBTQIA2S+ individual, activist or ally. In addition, each annotator is filtered by HIT approval rate $\geq 93\%$, completed > 7,500 HITs and located within the United States.

CHAPTER 5

A MULTI-LAYERED LANGUAGE ANALYSIS: A CASE STUDY OF AFRICAN-AMERICAN ENGLISH

Currently, natural language processing (NLP) models proliferate language discrimination leading to potentially harmful societal impacts as a result of biased outcomes. For example, part-of-speech taggers trained on Mainstream American English (MAE) produce non-interpretable results when applied to African American English (AAE) as a result of language features not seen during training. In this work, we incorporate a human-in-the-loop paradigm to gain a better understanding of AAE speakers' behavior and their language use, and highlight the need for dialectal language inclusivity so that native AAE speakers can extensively interact with NLP systems while reducing feelings of disenfranchisement.

5.1 Introduction

Over the years, social media users have leveraged online conversational platforms to perpetually express themselves online. For example, African American English (AAE), an English language variety is often heavily used on Twitter [48, 13]. This dialect continuum is neither spoken by *all* African Americans or individuals who identify as BIPOC (Black, Indigenous, or People of Color), nor is it spoken *only* by African Americans or BIPOC individuals [48, 11]. In some cases, AAE, a low-resource language (LRL) may be the first (or dominant) language, rather than the second (or non-dominant) language of an English speaker.

Specifically, AAE is a regional dialect continuum that consists of a distinct set of lexical items, some of which have distinct semantic meanings, and may possess different syntactic structures/patterns than in Mainstream American English (MAE) (*e.g.*, differentiating habitual *be* and non-habitual *be* usage) [127, 45, 71, 48, 11, 6, 13, 79]. In particular, [56] states that AAE possesses a morphologically invariant form of the verb that distinguishes between habitual action and currently occurring action, namely *habitual* ***be***. For example, "the habitual be" experiment by University of Massachusetts Amherst's Janice Jackson.

However, AAE is perceived to be "bad english" despite numerous studies by socio/raciolinguists

Table 5.1: An illustrative example of POS tagging of semantically equivalent sentences written in MAE and linguistics features of AAE lexical items, and their misclassified NLTK (inferred) tags, respectively.

MAE	**Input**	I have never done this before
	Output	(I, <PRP>), (have, <VBP>), (never, <RB>), (done, <VBN>), (that, <IN>), (before, <IN>)
AAE	**Input**	I aint neva did dat befo
	Output	(I, <PRP>), (aint, <VBP>), (neva, <NN>), (did, <VBD>)(dat, <JJ>), (befo, <NN>)

and dialectologists in their attempts to quantify AAE as a legitimized language [6, 48, 11, 79].

> "*[T]he common misconception [is] that language use has primarily to do with words and what they mean. It doesn't. It has primarily to do with people and what* ***they*** *mean.*" – [30]

Recently, online AAE has influenced the generation of resources for AAE-like text for natural language (NLP) and corpus linguistic tasks *e.g.*, part-of-speech (POS) tagging [72, 16], language generation [57] and automatic speech recognition [45, 133]. POS tagging is a token-level text classification task where each token is assigned a corresponding word category label (see Table 5.1). It is an enabling tool for NLP applications such as a syntactic parsing, named entity recognition, corpus linguistics, etc. In this work, we incorporate a human-in-the-loop paradigm by directly involving affected (user) communities to understand context and word ambiguities in an attempt to study dialectal language inclusivity in NLP language technologies that are generally designed for dominant language varieties. [34] state that,

> "*NLP systems aim to [learn] from natural language data, and mitigating social biases become a compelling matter not only for machine learning (ML) but for social justice as well.*"

To address these issues, we aim to empirically study *predictive bias* (see [129] for definition) *i.e.*, if POS tagger models make predictions dependent on demographic language features, and attempt a dynamic approach in data-collection of non-standard spellings and lexical items. To examine the behaviors of AAE speakers and their language use, we first collect variable (morphological and phonological) rules of AAE language features from literature [79, 4, 56, 11, 127, 14, 46, 6, 55] (see Appendix D.3). Then, we employ 5 trained sociolinguist Amazon Mechanical Turk (AMT)

annotators[1] who identify as bi-dialectal dominant AAE speakers to address the issue of lexical, semantic and syntactic ambiguity of tweets (see Appendix D.2 for annotation guidelines). Next, we incorporate a human-in-the-loop paradigm by recruiting 20 crowd-sourced diglossic annotators to evaluate AAE language variety (see Table 5.2). Finally, we conclude by expanding on the need for dialectal language inclusivity.

5.2 Related Work

Previous works regarding AAE linguistic features have analyzed tasks such as unsupervised domain adaptation for AAE-like language [72], detecting AAE syntax[127], language identification [15], voice recognition and transcription [45], dependency parsing [16], dialogue systems [85], hate speech/toxic language detection and examining racial bias [120, 58, 142, 38, 162, 102, 143, 78], and language generation [57]. These central works are conclusive for highlighting systematic biases of natural language processing (NLP) systems when employing AAE in common downstream tasks.

Although we mention popular works incorporating AAE, this dialectal continuum has been largely ignored and underrepresented by the NLP community in comparison to MAE. Such lack of language diversity cases constitutes technological inequality to minority groups, for example, by African Americans or BIPOC individuals, and may intensify feelings of disenfranchisement due to monolingualism. We refer to this pitfall as the *inconvenient truth i.e.*,

> "*[I]f the systems show discriminatory behaviors in the interactions, the user experience will be adversely affected.*" — [85]

Therefore, we define fairness as the model's ability to correctly predict each tag while performing zero-shot transfer via dialectal language inclusivity.

Moreover, these aforementioned works do not discuss nor reflect on the "*role of the speech and language technologies in sustaining language use*" [79, 10, 13] as,

> "*... models are expected to make predictions with the semantic information rather than with the demographic group identity information*" — [153].

[1] A HIT approval rate $\geq 95\%$ was used to select 5 bi-dialectal AMT annotators between the ages of 18 - 55, and completed > 10,000 HITs and located within the United States.

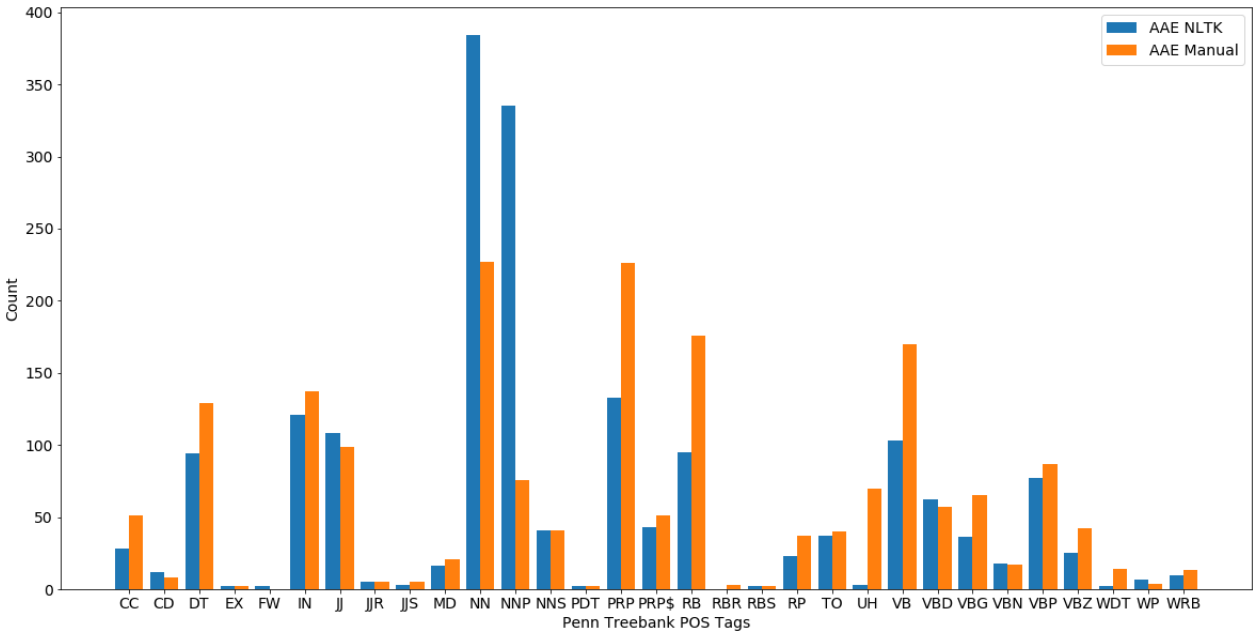

Figure 5.1: An illustration of inferred and manually-annotated AAE tag counts from k randomly sampled tweets.

Interactions with everyday items is increasingly mediated through language, yet systems have limited ability to process less-represented dialects such as AAE. For example, a common AAE phrase, "*I had a long* ***ass*** *day*" would receive a lower sentiment polarity score because of the word "***ass***", a (noun) term typically classified as offensive; however, in AAE, this term is often used as an emphatic, cumulative adjective and perceived as non-offensive.

Motivation: We want to test our hypothesis that training each model on correctly tagged AAE language features will improve the model's performance, interpretability, explainability, and usability to reduce predictive bias.

5.3 Dataset and Annotation

5.3.1 Dataset

We collect 3000 demographically-aligned African American (AA) tweets possessing an average of 7 words per tweet from the publicly available TwitterAAE corpus by [14]. Each tweet is accompanied by inferred geolocation topic model probabilities from Twitter + Census demographics

Table 5.2: Accurately tagged (observed) AAE and English phonological and morphological **linguistic** feature(s) accompanied by their respective MAE equivalent(s).

Tags	**Category**	**AAE Example(s)**	**MAE Equivalent(s)**
CC	Coordinating Conjunction	*doe/tho, n, bt*	*though, and, but*
DT	Determiner	*da, dis, dat*	*the, this, that*
EX	Existential There	*dea*	*there*
IN	Preposition/ Conjunction	*fa, cuz/cause, den*	*for, because, than*
JJ	Adjective	*foine, hawt*	*fine, hot*
PRP	Pronoun	*u, dey, dem*	*you, they, them*
PRP$	Personal Pronoun	*ha*	*her*
RB	Adverb	*tryna, finna, jus*	*trying to, fixing to, just*
RBR	Adverb, comparative	*mo, betta, hotta*	*more, better, hotter*
RP	Particle	*bout, thru*	*about, through*
TO	Infinite marker	*ta*	*to*
UH	Interjection	*wassup, ion, ian*	*what's up, I don't*
VBG	Verb, gerund	*sleepin, gettin*	*sleeping, getting*
VBZ	Verb, 3rd-person present tense	*iz*	*is*
WDT	Wh-determiner	*dat, wat, wus, wen*	*that, what, what's, when*
WRB	Wh-adverb	*hw*	*how*

and word likelihoods to calculate demographic dialect proportions. We aim to minimize (linguistic) discrimination by sampling tweets that possess over 99% confidence to develop "fair" NLP tools that are originally designed for dominant language varieties by integrating non-standardized varieties. More information about the TwitterAAE dataset, including its statistical information, annotation process, and the link(s) to downloadable versions can be found in Appendix D.1.

5.3.2 Preprocessing

As it is common for most words on social media to be plausibly semantically equivalent, we denoise each tweet as tweets typically possess unusual spelling patterns, repeated letter, emoticons and emojis[2]. We replace sequences of multiple repeated letters with three repeated letters (*e.g.*, *Hmmmmmmmm* $\rightarrow$ *Hmmm*), and remove all punctuation, "@" handles of users and emojis. Essentially, we aim to denoise each tweet only to capture non-standard spellings and lexical items more efficiently.

[2]Emoticons are particular textual features made of punctuation such as exclamation marks, letters, and/or numbers to create pictorial icons to display an emotion or sentiment (*e.g.*, ";)" $\Rightarrow$ *winking smile*), while emojis are small text-like pictographs of faces, objects, symbols, etc.

5.3.3 Annotation

First, we employ off-the-shelf taggers such as spacy and TwitterNLP; however, the Natural Language Toolkit (NLTK) [90] provides a more fine-grained Penn Treebank Tagset (PTB) along with evaluation metrics per tag such as F1 score. Next, we focus on aggregating the appropriate tags by collecting and manually-annotating tags from AAE/slang-specific dictionaries to assist the AMT annotators, and later we contrast these aggregated tags with inferred NLTK PTB inferred tags. In Figure 5.1, we display NLTK inferred and manually-annotated AAE tags from $k = 300$ randomly sampled tweets.

- **The Online Slang Dictionary** (American, English, and Urban slang) - created in 1996, this is the oldest web dictionary of slang words, neologisms, idioms, aphorisms, jargon, informal speech, and figurative usages. This dictionary possesses more than 24,000 real definitions and tags for over 17,000 slang words and phrases, 600 categories of meaning, word use mapping and aids in addressing lexical ambiguity.

- **Word Type** - an open source POS focused dictionary of words based on the Wiktionary project by Wikimedia. Researchers have parsed Wiktionary and other sources, including real definitions and categorical POS word use cases necessary to address the issue of lexical, semantic and syntactic ambiguity.

5.3.4 Human Evaluation

After an initial training of the AMT annotators, we task each annotator to annotate each tweet with the appropriate POS tags. Then, as a calibration study we attempt to measure the inter-annotator agreement (IAA) using Krippendorff's α. By using NLTK's [90] nltk.metrics.agreement, we calculate a Krippendorf's α of 0.88. We did not observe notable distinctions in annotator agreement across the individual tweets. We later randomly sampled 300 annotated tweets and recruit 20 crowd-sourced annotators to evaluate AAE language variety. To recruit 20 diglossic annotators[3],

[3]Note that we did not collect certain demographic information such as gender or race, only basic demographics such as age (18-55 years), state and country of residence.

we created a volunteer questionnaire with annotation guildlines, and released it on LinkedIn. The full annotation guildlines can be found in Appendix D.2. Each recruited annotator is tasked to judge sampled tweets and list their MAE equivalents to examine contextual differences of simple, deterministic morphosyntactic substitutions of dialect-specific vocabulary in standard English or MAE texts—a *reverse* study to highlight several varieties of AAE (see Table 5.2).

5.4 Methodology

In this section, we describe our approach to perform a preliminary study to validate the existence of predictive bias [46, 124] in POS models. We first introduce the POS tagging, and then propose two ML sequence models.

5.4.1 Part-of-Speech (POS) Tagging

We consider POS tagging as it represents word syntactic categories and serves as a pre-annotation tool for numerous downstream tasks, especially for non-standardized English language varieties such as AAE [151]. Common tags include prepositions, adjective, pronoun, noun, adverb, verb, interjection, etc., where multiple POS tags can be assigned to particular words due to syntactic structural patterns. This can also lead to misclassification of non-standardized words that do not exist in popular pre-trained NLP models.

5.4.2 Models

We propose to implement two well known sequence modeling algorithms, namely a Bidirectional Long Short Term Memory (Bi-LSTM) network, a deep neutral network (DNN) [63, 54] that has been used for POS tagging [84, 106], and a Conditional Random Field (CRF) [80] typically used to identify entities or patterns in texts by exploiting previously learned word data.

Taggers: First, we use NLTK [90] for automatic tagging; then, we pre-define a feature function for our CRF model where we optimized its L1 and L2 regularization parameters to 0.25 and 0.3, respectively. Later, we train our Bi-LSTM network for 40 epochs with an Adam optimizer, and a learning rate of 0.001. Note that each model would be accompanied by error analysis for a 70-30

split of the data with 5-fold cross-validation to obtain model classification reports, for metrics such as precision, recall and F1-score.

5.5 Operationalization of AAE as an English Language Variety

As (online) AAE can incorporate non-standardized spellings and lexical items, there is an active need for a human-in-the-loop paradigm as humans provide various forms of feedback in different stages of workflow. This can significantly improve the model's performance, interpretability, explainability, and usability. Therefore, crowd-sourcing to develop language technologies that consider who created the data will lead to the inclusion of diverse training data, and thus, decrease feelings of marginalization. For example, CORAAL, is an online resource that features AAL text data, recorded speech data, etc., into new and existing NLP technologies, AAE speakers can extensively interact with current NLP language technologies.

Consequently, to quantitatively and qualitatively ensure fairness in NLP tools, artificial intelligence (AI) and NLP researchers need to go beyond evaluation measures, word definitions and word order to assess AAE on a token-level to better understand context, culture and word ambiguities. We encourage both AI and NLP practitioners to prioritize collecting a set of relevant labeled training data with several examples of informal phrases, expressions, idioms, and regional-specific varieties. Specifically, in models intended for broad use such as sentiment analysis by partnering with low-resource and dialectal communities to develop impactful speech and language technologies for dialect continua such as AAE to minimize further stigmatization of an already stigmatized minority group.

5.6 Conclusion

Throughout this work, we highlight the need to develop language technologies for such varieties, pushing back against potentially discriminatory practices (in many cases, discriminatory through oversight more than malice). Our work calls for NLP researchers to consider both social and racial hierarchies sustained or intensified by current computational linguistic research. By shifting towards a human-in-the-loop paradigm to conduct deep multi-layered dialectal language analysis of AAE

to counter-attack erasure and several forms of biases such as *selection bias*, *label bias*, *model overamplification*, and *semantic bias* (see [124] for definitions) in NLP.

We hope our dynamic approach can encourage practitioners, researchers and developers for AAE inclusive work, and that our contributions can pave the way for normalizing the use of a human-in-the-loop paradigm both to obtain new data and create NLP tools to better comprehend underrepresented dialect continua and English language varieties. In this way, NLP community can revolutionize the ways in which humans and technology cooperate by considering certain demographic attributes such as culture, background, race and gender when developing and deploying NLP models.

5.7 Limitations And Ethical Considerations

All authors must warrant that increased model performance for non-standard varieties such as underrepresented dialects, non-standard spellings or lexical items in NLP systems can potentially enable automated discrimination. In this work, we *solely* attempt to highlight the need for dialectal inclusivity for the development of impactful speech and language technologies in the future, and do not intend for increased feelings of marginalization of an already stigmatized community.

CHAPTER 6

DETECTING AND MITIGATING INHERENT LINGUISTIC BIAS IN LARGE LANGUAGE MODELS

Recent studies show that NLP models trained on standard English texts tend to produce biased outcomes against underrepresented English varieties. In this work, we conduct a pioneering study of the English variety use of African American English (AAE) in NLI task. First, we propose CODESWITCH, a greedy unidirectional morphosyntactically-informed rule-based translation method for data augmentation. Next, we use CODESWITCH to present a preliminary study to determine if demographic language features do in fact influence models to produce false predictions. Then, we conduct experiments on two popular datasets and propose two simple, yet effective and generalizable debiasing methods. Our findings show that NLI models (e.g. BERT) trained under our proposed frameworks outperform traditional large language models while maintaining or even improving the prediction performance. In addition, we intend to release CODESWITCH, in hopes of promoting dialectal language diversity in training data to both reduce the discriminatory societal impacts and improve model robustness of downstream NLP tasks.

6.1 Introduction

In recent years, social media has become a pivotal tool its users to express their thoughts, feelings, and opinions on similar interests [37]. Typically, Standard American English (SAE), a high-resource language (HRL) is often used in formal communication, whereas African American English (AAE)[1] is primarily spoken in the United States and is often heavily and explicitly used on social media platforms such as Twitter [48, 13].

In particular, AAE is an English language variety and can be considered to be a low-resource language (LRL) that is neither spoken by *all* African Americans or individuals who identify as

[1]This English language variety has had several names within the last decades such as African American Vernacular English (AAVE), African American Language (AAL), Black English, Ebonics, Non-standard English, Northern Negro English and Black English Vernacular (BEV) [4, 56, 11, 76]. However, it is now commonly referred to as African American English (AAE), an English language variety.

BIPOC (Black, Indigenous, or People of Color), nor is it spoken *only* by African Americans or BIPOC individuals [48, 33, 11]. However, most dominant AAE speakers reside in diglossic communities and are able to *code-switch*, speaking both SAE and AAE. In linguistics, code-switching also referred to as language alternation is the ability of a speaker to alternate between two or more languages or language varieties within a particular conversation [149, 51, 40, 148, 33]. Thus, we refer to code-switching as switching among dialects, and/or language styles. For example, bi-dialectal AAE speakers are often able to code-switch between the SAE and both phonological and morphological language features of AAE while maintaining contextual intent.

Natural Language Understanding (NLU) is a subset of NLP, which enables human-computer interaction (HCI) by attempting to understand human language data such as text or speech, and communicate back to humans in their respective languages such as English, Spanish, etc., [121]. Hence, we will focus on *inference*, which is an eminent area of study of NLU. In particular, Natural language inference (NLI), a subset of NLU, also known as Recognizing Textual Entailment (RTE) is a segment-level categorization task of understanding the inferential relationships between sentence pairs and anticipating whether they are entailing, contradictory, or neutral sentences [23, 138].

Generally, the term *implicit bias* is used to refer to the unconscious preferential behaviors towards a certain demographic group such as age, race, ethnicity, gender, etc. [88, 131, 110]. However, in this study, to examine the differences in language styles from different demographic groups, we refer to this type of predisposed language style bias as *inherent linguistic bias*. Although, both biases are very similar, there exists a subtle difference as linguistic bias specifically refers to an analysis of every aspect of a particular language [161]. The existence of these biases in large language models (LLMs) such as mask language models (MLMs) generate language bias leading to potential harmful societal impacts inconveniencing members of LRL and diglossic communities who speak both standard languages and unrepresented dialects. This may increase feelings of marginalization and disenfranchisement [85, 13, 48].

Hence, in this work, we conduct a pioneering study of robustifying MLMs to minimize false predictions by introducing dialectal language diversity in training data to determine if MLMs learn

to make predictions based on demographic language features, and proposing two debias methods to enhance NLI models to mitigate the presence of linguistic bias during the training process. We posit that it is vital for production-ready MLMs improve their robustness to produce minimal systemic biases against protected attributes such as *race* and *gender* and thus, reducing discriminatory societal impacts [64, 126, 85, 131].

Specifically, we aim to answer two research questions: (1) *How can we as NLP practitioners encourage dialectal language diversity in training data?*; (2) *Do pretrained MLMs make predictions based on demographic language features?*; and (3) *How can we measure fairness and mitigate such biases in order to ensure fairness in NLU.*

Our contributions include:

- CODESWITCH, a greedy unidirectional morphosyntactically-informed rule-based translation method for data augmentation to generate intent-and-semantically equivalent AAE examples by perturbing SAE examples.
- Two intent-and-semantically equivalent NLI dataset of AAE sentence pairs with a wide range of morphological syntactic features and dialect-specific vocabulary.
- A detailed human evaluation of our human annotators to ensure contextual accuracy of adversarial sentence pairs (see Appendix E.4 for details).
- Two simple, yet effective debiasing methods to mitigate the inherent linguistic bias in NLI models, while maintaining or even improving their prediction performance.

6.2 Preliminaries

In this section, we introduce some preliminary knowledge about the problem under study. We first present the problem statement, and then describe two popular NLI datasets used in our research.

6.2.1 Problem Statement

We aim to investigate sentence representations of two linguistic systems of different demographic groups to demonstrate the existence of constitutional linguistic bias. To address the above research

Table 6.1: Randomly chosen original SNLI and MNLI examples and their inferential relationships.

Dataset	Premise	Hypothesis	Label
SNLI	A land rover is being driven across a river.	A vehicle is crossing a river.	entailment
	Children smiling and waving at camera	They are smiling at their parents	neutral
	An older man is drinking orange juice at a restaurant.	Two women are at a restaurant drinking wine.	contradiction
MNLI	So i have to find a way to supplement that	I need a way to add something extra.	entailment
	The new rights are nice enough	Everyone really likes the newest benefits	neutral
	I don't know um do you do a lot of camping	I know exactly.	contradiction

questions, we define two goals:

1. The first goal is to predict inferential relationships between paired sentences i.e., the second sentence is an entailment, contradiction, or neutral with respect to the first sentence.

2. The second goal is to debias the sentence representations obtained from the words in the given sentence. Specifically, we want the sentence representation to *only* include the semantic information, but not the language style, whether SAE or AAE. Therefore, we want the MLM to ignore the language style of each demographic group in order to make fair predictions.

Mitigating such linguistic biases can help develop robust MLMs for LRLs and dialectal languages more easily. Our main objective is to focus on dialectal language inclusivity, while using the benefit of large pretrained MLMs in order to improve model robustness of downstream tasks of NLP technologies for LRLs and language varieties.

6.2.2 Dataset

In this subsection, we introduce two of the largest, most popular NLP datasets for textual inference, namely, the Stanford Natural Language Inference (SNLI) and Multi-Genre Natural Language Inference (MNLI) corpora.

6.2.2.1 SNLI corpus

The SNLI [23] corpus is constructed from the Flickr30k corpus [147]. The original image caption is classified as the *premise*, whereas, the *hypothesis* is a human-written *premise*-related sentence that must satisfy one of one of three relational conditions: (1) *Entailment* – true image description, (2) *Neutral* – neutral image description, and (3) *Contradiction* – false or random image

Table 6.2: Augmented SNLI and MNLI examples (from Table 6.1) following the application of CODESWITCH. Each blue highlight corresponds to the AAE equivalent from their respective SAE counterpart.

Dataset	Premise	Hypothesis	Label
SNLI AAE	A land rover ***bein*** driven across a river.	A vehicle ***crossin*** a river.	entailment
	Children ***smilin n wavin*** at camera	***Dey smilin*** at ***they*** parents	neutral
	A older man ***drinkin*** orange juice at a restaurant.	Two women at a restaurant ***drinkin*** wine.	contradiction
MNLI AAE	So i ***gotta*** find a way ***ta*** supplement ***dat***	I need a way ***ta*** add ***sumn*** extra.	entailment
	Da new rights nice enough	***Everybody*** really likes ***da*** newest benefits	neutral
	Ion kno um do ***u*** do a lot of ***campin***	I ***kno*** exactly.	contradiction

description. The SNLI corpus is a collection of 570K *premise-hypothesis* sentence pairs, where each pair is aligned with one of these three relational labels.

6.2.2.2 MNLI corpus

Similarly to SNLI, the MNLI corpus [138] is a closely related crowd-sourced collection of 433k sentence pairs and their relational labels. However, MNLI contains 10 distinct genre categories (i.e., *Letters, Verbatim, Fiction, Face-to-face, Travel, Telephone, Travel, Oxford University Press, Slate, 9/11,* and *Government*) written and spoken data instead of image caption data.

6.3 CODESWITCH Creation

In this section, we first describe the process of the creation of CODESWITCH, carried out in three steps: 1) data collection of morphological syntactic features and dialect-specific vocabulary, 2) candidate retrieval of simple, deterministic morphosyntactic substitutions for unidirectional translations, and 3) human evaluation to test contextual accuracy of perturbations generated by CODESWITCH.

6.3.1 Data Collection

First, to gain an better understanding of AAE language, we engage with literature, sample text examples and mass collect morpho-syntax rules (which we adapt from the literature) (see Appendix E.2) [4, 56, 11, 33, 13, 127, 14, 46]. Therefore, we attempt a proactive approach in data-collection of grammatical, structural and syntactic rules of word case usage of AAE language features to understand the application of AAE in NLP downstream tasks. Next, we employ and assist 6 trained

sociolinguist Amazon Mechanical Turk (AMT) workers[2] with our collected set rules and text examples.

Pairwise Sample Collection We first randomly sample n = 5000 SAE premise-hypothesis sentence pairs that contain at least 8 words from both SNLI and MNLI corpora for a total of 10,000 sentence pairs. For contextual accuracy, we task the first 3 workers to obtain the AAE equivalents of our SAE samples (see Table 6.1), where each annotator is tasked to translate each SAE sentence pair into AAE. The full annotation guidelines can be seen in Appendix E.3.

6.3.2 Candidate Retrieval

Starting from data collection, we next retrieve candidate phrases and words use cases for data augmentation from our obtained AAE equivalent sentence pairs. As [88] uses a deep text classification model to illustrate that demographic language features do in fact influence models to produce false predictions on semantically equivalent SAE and AAE texts, our protocol follows simple, deterministic substitutions of English texts by dialect-specific vocabulary. To do so, we make use of both SAE and AAE sentence pairs in a pairwise fashion and construct a unidirectional informed-based translative morpho-syntax protocol (TMsP) that enables CODESWITCH to convert any given SAE text to a text possessing adequate language features to be considered as AAE from a dominant AAE speaker. More details on TMsP can be found in Appendix E.2).

Obtaining new texts for downstream tasks from authors of certain demographic groups is time-consuming and requires heavy human labor [88, 33]. Therefore, we create CODESWITCH (see Algorithm 1), a greedy unidirectional morphosyntactically-informed rule-based translation method which is not only fast, but also functions as a human-in-the-loop paradigm; therefore, drastically reduces heavy human labor. Our approach for intent-and-semantically equivalent AAE data augmentation is intuitively simple and effective. Consequently, we can now explore code-switching in several NLP tasks to determine if LLMs such as MLMs learn to make predictions

[2]Each AMT worker is independent and a trained sociolinguist filtered by HIT approval rate $\geq$ 96%, completed > 10,000 HITs and location (within the United States)

Algorithm 1 The translative syntactic morphological method for CODESWITCH.

Input: Original SAE sequence x
Output: Translated AAE sequence x'
begin function
Load SAE input sequence $\rightarrow x$
$x \leftarrow$ LOWER(x)
$T \leftarrow$ TOKENIZE(x)
for all $i = 1, 2, ..., |T|$ **do**
  if $i \in$ {TMsP} **then**
    $T_i \leftarrow$CODESWITCH(i)
  end if
end for
$x' \leftarrow$ DETOKENIZE(T)
return x'
end function

based on demographic/ dialectal language features.

We represent each original NLI corpus as $D < P, H, L >$ with $p \in P$ as the premise, $h \in H$ as the hypothesis and, lastly, $l \in L$ as the label, and create two augmented datasets i.e., SNLI AAE and MNLI AAE, where we represent each augmented NLI dataset as $D' < P', H', L >$. Specifically, translate each premise-hypothesis pair to AAE and keep the original label unchanged to form a new instance. It is important to note that the task of CODESWITCH is to ensure both sets of datasets i.e., D and D' maintain their contextual accuracy, although they consist of two different language styles (see Table 6.2).

6.3.3 Human Evaluation

After an initial training of the AMT annotators with our annotation guidelines, we implement a minor calibration study by tasking the remaining 3 independent workers to test our AAE data augmentation method. We randomly sample 200 SAE/AAE sentence pair examples from each of the 4 datasets, for a total of 800 sentence pairs (or 1600 SAE/AAE sentences). The workers were asked to indicate (1) whether the AAE sentences are written by an L1 (or dominant) AAE speaker, or most likely to be machine generated (MG); and (2) whether or not their contextual accuracy is maintained. For content analysis to ensure the quality of our AAE samples and to quantify the extent of agreement between raters, we first let 3 annotators independently rate each AAE-generated

sentence pair as "Native" or "MG", then we measure the inter-annotator agreement (IAA) using Krippendorff's α.

We calculate an inter-rater reliability of 0.82, and did not observe significant differences in agreement across the individual sentences. Qualitative analysis revealed that generated samples resembled sequences written by L1 AAE speakers, whereas few samples were classified as most likely MG. Annotators informed us of particular morpho-syntax cases, for example, constant copula deletion of the verb "**be**" and its variants, namely "**is**" and "**are**" is irregular and often inserted last in word order. This indicates that CODESWITCH does not account for contextual instances when generating AAE samples, hence being classified as most likely MG.

6.4 Empirical Study and Analysis

In this section, we conduct a preliminary study to substantiate the existence of inherent linguistic bias in NLI models. We introduce the base NLI models and training details, and then we demonstrate our empirical results.

To illustrate inherent linguistic bias of two distinct linguistic systems, we introduce a representative MLM, namely, BERT [41] (see Appendix E.1 for more details).

Table 6.3: Model performance when tested on AAE data.

	Model Performance (%)					
	SNLI			**MNLI**		
Models	SAE	AAE	Diff.	SAE	AAE	Diff.
$\text{BERT}_{\text{BASE}}$	90.12	86	-4.12	84.77	79.79	-4.68
$\text{BERT}_{\text{LARGE}}$	90.46	74.55	-15.91	84.47	67.35	-17.12

We use each original dataset i.e., SNLI and MNLI to fine-tune both BERT models on a batch size of 32 using an AdamW optimizer with a learning rate of 2e-5 and default betas (β_1 = 0.9, β_2 = 0.999) for 3 epochs. Our experiments display that pretrained MLMs "*are only as good as the data they are trained on*" and are unable to make fair predictions [131]. In Table 6.3, we see that the lack of diverse training data results in disparities in model performance in MLMs, which may be significantly be intensified as models become more complex. In Table 6.4, we illustrate several examples on the inherent linguistic bias on account of demographic language features, and can

Table 6.4: An illustration of inherent linguistic bias between AAE and their respective SAE counterpart (see Appendix E.2).

Premise	Hypothesis	Label	Prediction
Dis church choir sings ***ta da*** masses as ***dey*** sing joyous songs from ***da*** book at a church.	***Da*** church filled ***wit*** song.	Entailment	Neutral
Dis church choir sings ***ta da*** masses as ***dey*** sing joyous songs from ***da*** book at a church.	***Da*** church has cracks in ***da*** ceiling.	Neutral	Contradiction
Dis church choir sings ***ta da*** masses as ***dey*** sing joyous songs from ***da*** book at a church.	A choir ***singin*** at a baseball game.	Contradiction	Entailment
A woman ***wit*** a green headscarf, blue shirt***n*** a very big grin.	***Da*** woman young.	Neutral	Contradiction
A woman ***wit*** a green headscarf, blue shirt ***n*** a very big grin.	***Da*** woman very happy.	Entailment	Neutral

conclude that demographic/ dialectal language features do in fact influence models to produce false predictions.

6.5 Debiasing Methods

In Section 6.4, we empirically demonstrate that popular NLI models show significant bias towards AAE by underperforming on them than SAE. A natural question arises: *how can we remove the biases in NLI models towards different language styles?* To solve this problem, we introduce two simple but effective debiasing strategies: (1) counterpart data augmentation (CDA); and (2) language Style disentanglement (LSD).

6.5.1 Counterpart Data Augmentation

The bias of NLI models originates from the training data. Since the training data contains only SAE, the NLI models trained on such data does not understand the unique vocabulary and grammar of AAE, which leads to poor performance. Thus, we propose to implement CODESWITCH to augment the original SAE training data by translating them to their AAE counterparts and in turn implement CDA strategy similar to [158, 163]. Then, we will get a large augmented training dataset, D^+, which is twice the size of the original datasets (*i.e.,* SNLI) as it contains both D and D'.

6.5.2 Language Style Disentanglement

For two texts with the similar intent and semantic content of different language styles (e.g. SAE v.s. AAE), an NLI model may tend to make biased predictions towards one style. The

immediate reason is that the NLI prediction are based on the language style features, instead of relying solely on the semantic features of the texts. Based on this consideration, we propose LSD, an in-processing debiasing method, which tries to disentangle the language style features from the semantic features in text representations and forces the NLI model to make inference on the pure semantic representations.

6.5.2.1 The LSD Framework

To achieve disentanglement, we adopt the idea of adversarial learning. Figure 6.1 illustrates the overall framework of LSD. We view the framework as three parts: (1) the BERT model that encodes a premise-hypothesis pair as a fixed-dimensional representation $\mathbf{E}_{[CLS]}$; (2) a feed-forward neural (FFN) classifier $\mathbf{C}$ that takes $\mathbf{E}_{[CLS]}$ as input to predict the inferential relationship between the premise and the hypothesis; and (3) a FFN discriminator $\mathbf{D}$ that predicts whether the sentence pair is SAE or AAE based on $\mathbf{E}_{[CLS]}$. Via adversarial learning, our goal is to build a BERT model that can produce an accurate semantic representation of the text pair so that the classifier $\mathbf{C}$ can make correct predictions based on it, while the representation is free from the language style features of the texts, so that the discriminator $\mathbf{D}$ cannot distinguish whether the texts are from D or D'.

Algorithm 2 The optimization method for the LSD framework.

Input: Training data $\mathbf{T} = \{< P_i, H_i, L_i, S_i >\}_{i=1}^{|\mathbf{T}|}$ and Validation data $\mathbf{V} = \{< P_i, H_i, L_i, S_i >\}_{i=1}^{|\mathbf{V}|}$
Output: BERT parameters $\mathbf{W}^{\text{BERT}}$, classifier parameters $\mathbf{W}^{\mathbf{C}}$
Load pre-trained parameters $\mathbf{W}^{\text{BERT}}$
Initialize $\mathbf{W}^{\mathbf{C}}$ and $\mathbf{W}^{\mathbf{D}}$

for N epochs **do**
  for M batches **do**
    Obtain a mini-batch of training data $\mathbf{B}$ from $\mathbf{T}$
    Update $\mathbf{W}^{\mathbf{D}}$ by optimizing $L_{\mathbf{D}}$ in Equation 6.1
    Update $\mathbf{W}^{\text{BERT}}$ and $\mathbf{W}^{\mathbf{C}}$ by optimizing L in Equation 6.2
  end for
  Run the BERT model and the classifier $\mathbf{C}$ on validation data $\mathbf{V}$
  Save parameters $\mathbf{W}^{\text{BERT}}$ and $\mathbf{W}^{\mathbf{C}}$ if achieving the best validation performance so far.
end for

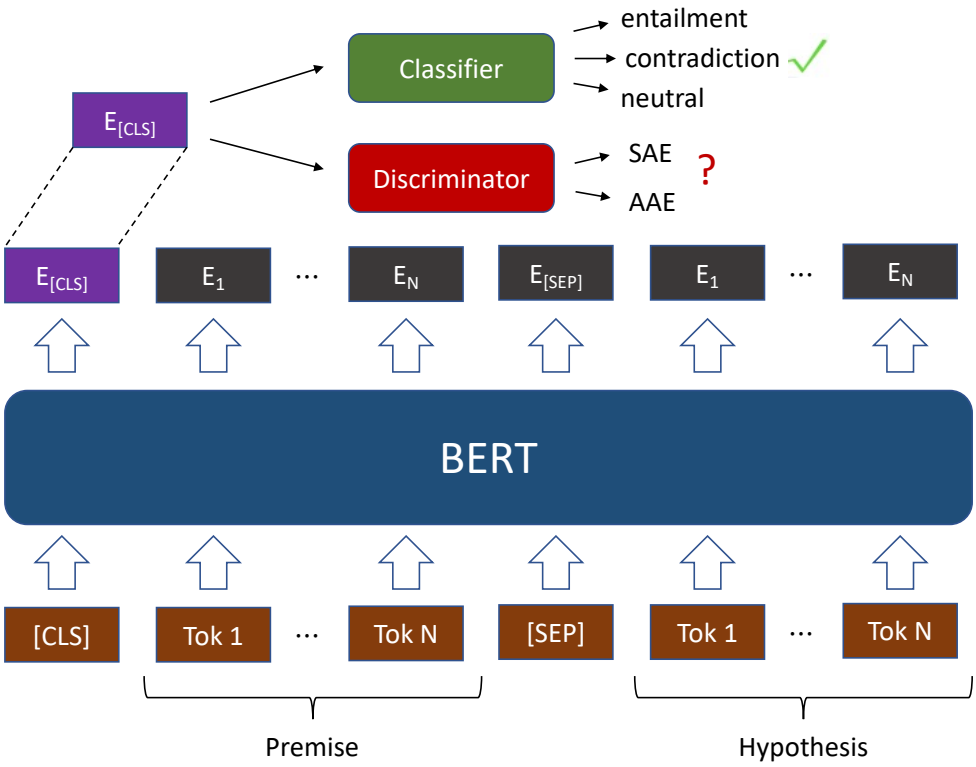

Figure 6.1: An illustration of the language-style disentanglement model.

6.5.2.2 An Optimization Method

We present our optimization algorithm for the LSD framework in Algorithm 2. We train the framework on the augmented training dataset obtained via our CODESWITCH method as we do in CDA. In the training data $\mathbf{T} = \{< P_i, H_i, L_i, S_i >\}_{i=1}^{|\mathbf{T}|}$, each instance consists of a premise p, a hypothesis h, a label l, and a binary language style label $S \in \{\text{SAE}, \text{AAE}\}$. At the beginning, we first load pretrained BERT parameters, and initialize the parameters of the classifier $\mathbf{C}$ and the discriminator $\mathbf{D}$ (line 3-4). In each iteration, we first obtain a mini-batch of training data $\mathbf{B} = \{< P_i, H_i, L_i, S_i >\}_{i=1}^{|\mathbf{B}|}$ (line 3). Then, we update the discriminator $\mathbf{D}$ by minimizing the following cross-entropy loss (line 4):

$$L_{\mathbf{D}} = -(\mathbb{I}\{S = 0\} \log p_0^{\mathbf{D}} + \mathbb{I}\{S = 1\} \log p_1^{\mathbf{D}}) \tag{6.1}$$

where S is the language style label of the utterance. $S = 0$ represents for SAE and $S = 1$ represents for AAE. $p_0^{\mathbf{D}}$ and $p_1^{\mathbf{D}}$ are the two elements in the predicted probability $\mathbf{p}^{\mathbf{D}}$ from the discriminator $\mathbf{D}$.

Minimizing $L_{\mathbf{D}}$ will force **D** to make correct predictions.

Next, we calculate the cross-entropy loss on the main prediction task:

$$L_{\mathbf{C}} = -(\mathbb{I}\{L=0\}\log p_0^{\mathbf{C}} + \mathbb{I}\{L=1\}\log p_1^{\mathbf{C}} + \mathbb{I}\{L=2\}\log p_2^{\mathbf{C}})$$

where L is the set of labels of the NLI task. $S = 0, 1, 2$ represent for entailment, contradiction, and neutral, respectively. $p_j^{\mathbf{C}}$ indicates the predicted probability for the j-th label from the classifier **C**. Minimizing $L_{\mathbf{C}}$ will force **C** to make correct predictions. To ensure that the BERT model produces a text representation that can fool the discriminator, when training, we consider another entropy loss:

$$L_{\mathbf{D}'} = -(p_0^{\mathbf{D}}\log p_0^{\mathbf{D}} + p_1^{\mathbf{D}}\log p_1^{\mathbf{D}})$$

$L_{\mathbf{D}'}$ is the entropy of the predicted distribution $\mathbf{p}^{\mathbf{D}}$ from the discriminator. Minimizing it makes $\mathbf{p}^{\mathbf{D}}$ close to an even distribution, preventing **D** from making correct predictions. We update the BERT model and the classifier by minimizing the following combined loss (line 5):

$$L = L_{\mathbf{C}} + L_{\mathbf{D}'} \tag{6.2}$$

At the end of each epoch, we run the BERT model and the classifier on the validation data, and save their parameters if they achieve the best validation performance.

6.5.3 Experimental results

In Table 6.5, we show the performances of the two debiasing methods on two datasets in terms of two BERT models. In Table 6.3, the results of the debiased models CDA, LSD and that of the original models were compared. Note that our two debiasing methods reduce the gap between the performances on SAE and AAE significantly. The original BERT models perform well on SAE test data but exhibit a decrease in performance when they are tested on AAE data. However, the BERT models trained under CDA or LSD debiasing strategies achieve similar model performance on SAE and AAE, which demonstrates the effectiveness of the two debiasing methods to mitigate bias in NLI models.

Table 6.5: Model performances of two debiased NLI models.

	Model Performance (%)					
	SNLI			MultiNLI		
Models	SAE	AAE	Diff.	SAE	AAE	Diff.
CDA_{BASE}	89.77	89.76	-0.01	84.29	83.98	-0.31
LSD_{BASE}	90.35	90.49	+0.14	84.50	83.81	-0.69
CDA_{LARGE}	90.48	90.36	-0.12	84.66	84.20	-0.46
LSD_{LARGE}	90.60	90.53	-0.07	84.72	84.30	-0.42

Furthermore, our debiased models not only improve the performance on AAE data, but also maintain similar performance on SAE data as the original model. This is due to either the introduction of additional AAE training data which is not always available, and the disentanglement between the semantic and language style features of texts enhancing the model's capability of understanding natural language. Lastly, we find that LSD generally outperforms CDA on both SAE and AAE data. In addition, LSD is an adversarial learning debaising method that filters out irrelevant language style information towards the NLI task. In fact, LSD is also generalizable for more effective and architecturally similar models such as DeBERTa [61], XLNet [144], and T5 [109] to ensure fairness as well as robustifying larger language models.

6.6 Related Work

Previous works focus on AAE in the context of *racial bias* as a result of systemic biases in model performance. For example, [16] focus on dependency parsing social media AAE to analyze the impacts of performance disparities between AAE and SAE tweets. Other works undertake AAE within the scope of detecting and mitigating the presence of racial bias in areas of offensive and abusive language detection [85, 120], sentiment analysis [57] and hate speech detection [39, 120]. However, these influential works do not engage with AAE literature, utilize a human-in-the-loop paradigm nor employ the humans who create such data. Thus, these pivotal works fail to understand AAE's phonological and morphological language features—thereby simply treating AAE as another non-Penn Treebank English variety [13].

Fairness in NLP. As social and racial disparities have become a compelling issue within the NLP community, focal topics of fairness, accountability, ethics, sustainable development, etc., have

gained momentous attention in recent years [64]. Recent work on fairness has primarily been focused on racial and gender biases in distributed word representations [17, 158, 163], coreference resolution [117], sentence encoders [95], machine translation [132, 107], and dialogue generation [85, 89].

Adversarial learning in NLP. Adversarial examples were initially explored in computer vision by [130], where these examples were intended to influence models to produce false predictions. However, in NLP, adversarial examples can occur at a phonetic, phonological, morphological, syntactic, semantic, or pragmatic level [131, 40, 51, 149]. [85] displays that dialogue systems are prone to produce offensive responses when fed AAE language features in comparison to SAE, whereas [89] propose a novel adversarial learning framework which directly addresses the issue of gender bias in dialogue models while maintaining their performance. Both [1] and [73] exploit the notion of adversariality by utilizing word embeddings to find the k nearest synonymic examples.

Summary. These influential works demonstrate novel adversarial learning methodologies on a character and/or word-level in order to address bias issues surrounding protected attributes such as race and gender by improving model robustness. Similarly, our work utilizes a human-in-the-loop paradigm by employing humans who create such data, to create a novel morphosyntactic method to perturb language styles on a syntactic-level to highlight the need for dialectal language diversity in training data.

6.7 Conclusion and Future Works

To address compelling fairness, accountability, transparency, and ethical concerns surrounding the sustainability of language use in NLP applications, we claim that the addition of diverse dialectal language in training data will improve model robustness and generalizability. Our findings show that our proposed debiasing methods not only improves the performance on AAE data but effectively reduces the performance gap between SAE and AAE significantly, while maintaining or even improving the prediction performance on SAE data. Therefore, training under these two debiasing strategies aids in the mitigation of linguistic bias in NLI models.

We conclude that though similar, the two language styles, SAE and AAE are not identical,

and thus, should not solely be evaluated against each other, but compared to as a basis of model performance minimize the existence of inherent linguistic bias in language models. In the future, we intend to release CODESWITCH a morphosyntactically-informed rule-based translation method for unidirectional data augmentation for generating intent-and-semantically-equivalent AAE examples as a public python package, to encourage further computational linguistic research into debiasing various NLP systems. We actively intend on updating CODESWITCH s.t. it can include new or regional-specific *lingo*. In this way, CODESWITCH can constitute potential groundwork on ways that AAE can effectively be integrated in NLP systems to improve future language models during their development and employment.

6.8 Limitations And Ethical Considerations

All authors must warrant mentioning that the increased performance for underrepresented dialects in NLP systems has the potential to enable automated discrimination based on the use of non-standard dialects. Although, we attempt to highlight the need for dialectal inclusivity for impactful speech and language technologies, we do not intend for increased feelings of marginalization of an already stigmatized community.

We have established our method's effectiveness for data augmentation for generating intent-and-semantically-equivalent AAE examples and believe that CODESWITCH could be further improved by addressing the following limitations:

1. Currently, CODESWITCH is a unidirectional data augmentation method and cannot be used in reverse as a deterministic text normalization/preprocessing system which can convert all text to SAE.

2. CODESWITCH operates on simple, deterministic substitutions for morphosyntactically-informed translations rules found in Appendix E.2 rather than that of real L1 and L2 AAE speakers, which may result in the lack of several formal/informal phrases, expressions, idioms, cultural and regional-specific lingo, and slang-related words [13]. For example, "I ***sholl*** was finna ask who money dat is ", where "*sholl*" refer to the replacement of the word "*sure*".

3. Although CODESWITCH possesses several simple, deterministic morphosyntactically-informed translation rules it does account for contextual instances of accurate copula deletion. This may lead to a discrepancy between actual text written by L1 and/or L2 AAE speakers and our proposed data augmentation method.

In the future, we intend to address these limitations and ethical considerations by partnering with AAE diglossic communities in hopes of robustifying CODESWITCH to be probabilistic rather than deterministic to capture different AAE variants of the same SAE term (for example, the AAE equivalents to "what's" → "*waz*" or "*wus*" or "*wats*". In addition, we will investigate inherent linguistic bias in other NLP applications.

CHAPTER 7

CONCLUSION

7.1 Dissertation Summary

First, in Chapter 2, we introduce our first case study, namely, *Gender, Race, Language and Social Justice*, we conduct a pioneering study about the fairness issues concerning both gender and racial biases in two popular dialogue models, i.e., generative and retrieval dialogue models as a joint problem. We detect and demonstrate performance disparities between sequence generation between gender (male/female) and racial (white/black) responses, respectively. To address this aforementioned issue, we propose two simple but effective debiasing methods to reduce these disparities and to better facilitate issues surrounding social justice and fairness in dialogue generation.

Next, in the case study, *Gender and Sexual Identities, Orientations and Expressions*, we examine predictive biases of both binary (male and female) representations in Chapter 3 and LGBTQIA2S+ representations in Chapter 4, respectfully. In Chapter 3, we construct two large benchmark datasets: (1) possessive (gender-specific and gender-neutral) nouns dataset and (2) attribute (career-related and family-related) words dataset to systematically conduct large scale analyses of each news corpora to detect and examine gender biases in distribution, content, and labeling and word choice, and demonstrate that societal gender biases in regards to gender roles in society. Moreover, we learn that gender is a progressive spectrum attempting to include all gender and sexual identities and orientations. However, individuals who may oppose, criticize, or possess contradictory feelings, beliefs, or motivations towards certain communities constitute discrimination, harassment, and abuse in the form of hate speech, abusive and offensive language use [35, 37, 12, 8]. To address the above challenges, in Chapter 4, we aim to intersect NLP, gender and queerness to readily identify and detect online sexist and Anti-LGBTQIA2S+ content. To the best of our knowledge, our dataset is first dataset for scientists, practitioners and researchers to study stereotypical social biases against this already marginalized community in hopes towards inclusivity.

Later, in the case study, *Language, Race and Culture* is divided into two folds. In Chapter

5, we incorporate a human-in-the-loop paradigm to address the issue of lexical, semantic and syntactic ambiguity of African American English (AAE) use in traditional off-the-shelf models and well-known large language models (LLMs). We further propose a dataset of tweets containing labeled AAE morphosyntactic lexical features in order to enable sociolinguistic, raciolinguistics and dialectologists analysis of morphosyntactic variation in AAE. We provide a normative foundation for reasoning about harms arising from NLP systems that we have shown is largely absent from the current literature. In Chapter 6, we conduct a pioneering study of robustifying large language models (LLMs) to minimize false predictions (or error disparities) by introducing dialectal language to diversify training data to provide a normative foundation for reasoning about cultural and linguistic harms arising from state-of-the art NLP systems. To substantiate the existence of inherent linguistic bias in LLMs, we attempt a dynamic approach to generate an AAE sentence pair dataset with a wide range of morphological syntactic features and dialect-specific vocabulary to illustrate that the lack of diverse training data results in disparities in model performance. To do so, we construct CodeSwitch, a greedy unidirectional morphosyntactic translation method for data augmentation to generate intent-and-semantically equivalent AAE examples by perturbing SAE examples.

Finally, in Chapter 7 concludes our works and we identify possible future directions drawing on work across Trustworthy AI and its applications, in an attempt to provide a foundation through an account of the relationships between language and injustice to develop a unified view and to build AI tools that combine the best characteristics of all.

7.2 Future Work

I am extremely thrilled by the potential of my research area for current and future technology. The ultimate goal is to develop a unified view for all and to build tools that combine the best characteristics of all. Moreover, I plan to extend the scope of incremental pattern discovery framework in various directions, for example, creating safe and trustworthy AI technologies for people with disabilities (PWDs).

In the near future, I plan to work with faculty, students, research scientists as my long-term career goal is to develop advanced technologies for various applications of FATE in AI, ML and

NLP to develop Trustworthy AI technology. To ensure that these research advances and innovations have a positive impact on society I intend to continue my inclusive approach of directly involving affected (user) communities to address social issues surrounding social biases (e.g. *gender* and *racial* biases). For example, tackling several problems such as the gender gap in facial recognition systems and facial recognition disparities across demographics, understanding dialect disparities in Natural Language Understanding (NLU), disambiguation of morphosyntactic features of AAE–the case of habitual "*be*", improving genetic risk prediction across diverse population by disentangling ancestry representations, analyzing hate speech data along a racial, gender and intersectional axes, misinformation, etc. In the futher future, I aim to explore offensive language mitigation solutions towards members of the LGBTQIA2S+ community with the objective of creating a safe and inclusive place that welcomes, supports, and values all LGBTQIA2S+ individuals, activists and allies both online and offline.

7.3 Concluding Remarks

The goal of NLP is to process language at a human level. However, NLP's current approach-ignoring social factors-prevents us from reaching human-level competence and performance since language is more than just informational content. As such, I will continuously evaluate my research, collaborating with technical and non-technical audiences to gain new perspectives, and challenge myself to improve daily. I will also maintain active interest in related research areas, from which I derive a rich supply of ideas and techniques to tackle new and existing problems. By working at the edges between theory and practice, I hope to make unique and lasting contributions to the social and scientific communities as I believe to create new and innovative technology tomorrow, we need to start today.

BIBLIOGRAPHY

[1] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2890–2896, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[2] Mohadeseh Amini and Parviz Birjandi. Gender bias in the iranian high school efl textbooks. English Language Teaching, 5(2):134–147, 2012.

[3] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, pages 405–413, 2019.

[4] Guy Bailey, John Baughan, Salikoko S. Mufwene, and John R. Rickford. African-American English: Structure, History and Use (1st ed.). Routledge, 1998.

[5] Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, August 2021. Association for Computational Linguistics.

[6] John Baugh. Linguistic discrimination. In 1. Halbband, pages 709–714. De Gruyter Mouton, 2008.

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. CoRR, abs/1706.02409, 2017.

[8] Federico Bianchi and Dirk Hovy. On the gap between adoption and understanding in NLP. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3895–3901, Online, August 2021. Association for Computational Linguistics.

[9] Steven Bird. NLTK: the natural language toolkit. In ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006, 2006.

[10] Steven Bird. Decolonising speech and language technology. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3504–3519, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[11] Linda M. Bland-Stewart. Difference or deficit in speakers of african american english? https://leader.pubs.asha.org/doi/10.1044/leader.FTR1.10062005.6, May 2005.

[12] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of “bias” in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5454–5476, Online, July 2020. Association for Computational Linguistics.

[13] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. Language (technology) is power: A critical survey of "bias" in NLP. CoRR, abs/2005.14050, 2020.

[14] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics.

[15] Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media african-american english. CoRR, abs/1707.00061, 2017.

[16] 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), pages 1415–1425, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[17] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, page 4356–4364, 2016.

[18] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 4349–4357. Curran Associates, Inc., 2016.

[19] Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models. CoRR, abs/1904.03035, 2019.

[20] Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models, 2019.

[21] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In Companion Proceedings of The 2019 World Wide Web Conference, pages 491–500, 2019.

[22] Avishek Joey Bose and William Hamilton. Compositional fairness constraints for graph embeddings. CoRR, abs/1905.10674, 2019.

[23] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large

annotated corpus for learning natural language inference. CoRR, abs/1508.05326, 2015.

[24] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. American Association for the Advancement of Science, 2017.

[25] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. 356:183–186, April 2017.

[26] Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Philip McCrae. Dataset for identification of homophobia and transophobia in multilingual youtube comments. CoRR, abs/2109.00227, 2021.

[27] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. CoRR, abs/1711.01731, 2017.

[28] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. Proportionally fair clustering. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, pages 1032–1041, 2019.

[29] Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. On measuring gender bias in translation of gender-neutral pronouns, 2019.

[30] Herbert H. Clark and Michael F. Schober. Asking questions and influencing answers. In Russell Sage Foundation, 1992.

[31] Steven R. Corman, Timothy Kuhn, Robert D. Mcphee, and Kevin J. Dooley. Studying complex discursive systems. Human Communication Research.

[32] Jamell Dacon. Recommender system datasets. https://github.com/daconjam/Recommender-System-Datasets, 2020.

[33] Jamell Dacon. Towards a deep multi-layered dialectal language analysis: A case study of African-American English. In Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 55–63, Seattle, Washington, July 2022. Association for Computational Linguistics.

[34] Jamell Dacon and Haochen Liu. Does gender matter in the news? detecting and examining gender bias in news articles. In Companion Proceedings of the Web Conference 2021, WWW '21, page 385–392, New York, NY, USA, 2021. Association for Computing Machinery.

[35] Jamell Dacon and Haochen Liu. Does gender matter in the news? detecting and examining gender bias in news articles. New York, NY, USA, 2021. Association for Computing Machinery.

[36] Jamell Dacon, Harry Shomer, Shaylynn Crum-Dacon, and Jiliang Tang. Detecting harmful online conversational content towards lgbtqia+ individuals. In Queer in AI Workshop at NAACL, 2022.

[37] Jamell Dacon and Jiliang Tang. What truly matters? using linguistic cues for analyzing the #blacklivesmatter movement and its counter protests: 2013 to 2020. CoRR, abs/2109.12192, 2021.

[38] Thomas Davidson and Debasmita Bhattacharya. Examining racial bias in an online abuse corpus with structural topic modeling. CoRR, abs/2005.13041, 2020.

[39] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In Proceedings of the Third Workshop on Abusive Language Online, pages 25–35, Florence, Italy, August 2019. Association for Computational Linguistics.

[40] Charles E. DeBose. Codeswitching: Black english and standard english in the african-american linguistic repertoire. Journal of Multilingual and Multicultural Development, 13(1-2):157–167, 1992.

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018.

[42] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. Queens are powerful too: Mitigating gender bias in dialogue generation, 2020.

[43] Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. CoRR, abs/1908.06083, 2019.

[44] Jaschar Domann, Jens Meiners, Lea Helmers, and A. Lommatzsch. Real-time news recommendations using apache spark. In CLEF, 2016.

[45] Rachel Dorn. Dialect-specific models for automatic speech recognition of African American Vernacular English. In Proceedings of the Student Research Workshop Associated with RANLP 2019, pages 16–20, Varna, Bulgaria, September 2019. INCOMA Ltd.

[46] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 11–21, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[47] Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. In plain sight: Media bias through the lens of factual reporting. In EMNLP-IJCNLP, pages 6343–6349, Hong Kong, China, November 2019.

[48] Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. A survey of race, racism, and anti-racism in NLP. CoRR, abs/2106.11410, 2021.

[49] Joel Escudé Font and Marta R Costa-Jussa. Equalizing gender biases in neural machine translation with word embeddings techniques. arXiv preprint arXiv:1901.03116, 2019.

[50] Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to conversational AI. Foundations and Trends in Information Retrieval, 13(2-3):127–298, 2019.

[51] Penelope Gardner-Chloros et al. Code-switching. Cambridge university press, 2009.

[52] Danielle Gaucher, Justin Friesen, and Aaron C Kay. Evidence that gendered wording in job advertisements exists and sustains gender inequality. Journal of personality and social psychology, 101(1):109, 2011.

[53] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint arXiv:1903.03862, 2019.

[54] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks, 18(5):602–610, 2005. IJCNN 2005.

[55] Jonathon Green. The vulgar tongue: Green's history of slang. Oxford University Press, New York, USA, 2014.

[56] Lisa J. Green. African American English: A Linguistic Introduction. Cambridge University Press, 2002.

[57] Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. Investigating African-American Vernacular English in transformer-based text generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5877–5883, Online, November 2020. Association for Computational Linguistics.

[58] Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. New York, NY, USA, 2021. Association for Computing Machinery.

[59] Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: an interdisciplinary literature review. International Journal on Digital Libraries, pages 1–25, 2018.

[60] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

[61] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. CoRR, abs/2006.03654, 2020.

[62] Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. Ethical challenges in data-driven dialogue systems. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018, pages 123–129, 2018.

[63] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural Comput., 9(8):1735–1780, November 1997.

[64] Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 591–598, Berlin, Germany, August 2016. Association for Computational Linguistics.

[65] Ayanna Howard and Jason Borenstein. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and engineering ethics, 24(5):1521–1536, 2018.

[66] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In ICWSM, 2014.

[67] Clayton J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014., 2014.

[68] Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. CoRR, abs/1608.07187, 2016.

[69] Sen Jia, Thomas Lansdall-Welfare, and Nello Cristianini. Measuring gender bias in news images. In Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion. Association for Computing Machinery, 2015.

[70] Sen Jia, Thomas Lansdall-Welfare, Saatviga Sudhahar, Cynthia Carter, and Nello Cristianini. Women are seen more than heard in online newspapers. PloS one, 11:e0148434, 02 2016.

[71] Taylor Jones. Toward a description of african american vernacular english dialect regions using "black twitter". American Speech, 90:403–440, 11 2015.

[72] Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Learning a POS tagger for AAVE-like language. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages

1115–1120, San Diego, California, June 2016. Association for Computational Linguistics.

[73] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. CoRR, abs/1907.10529, 2019.

[74] Dan Jurafsky and James H. Martin. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, 2009.

[75] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Proceedings of the 2012th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II, ECMLPKDD'12, pages 35–50, Berlin, Heidelberg, 2012. Springer-Verlag.

[76] Sharese King. From african american vernacular english to african american language: Rethinking the study of race and language in african americans' speech. Annual Review of Linguistics, 6(1):285–300, 2020.

[77] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14):7684–7689, 2020.

[78] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14):7684–7689, 2020.

[79] William Labov. Ralph fasold, tense marking in black english: a linguistic and social analysis. washington, d.c.: Center for applied linguistics, 1972. pp. 254. Language in Society, 4(2):222–227, 1975.

[80] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[81] Andrienne Lafrance. I analyzed a year of my reporting for gender bias (again). https://www.theatlantic.com/technology/archive/2016/02/gender-diversity-journalism/463023/, February 2016.

[82] Xiaolan Lei. Sexism in language. Journal of Language and Linguistics, 5(1):87–94, 2006.

[83] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In NAACL HLT 2016, The

2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, pages 110–119, 2016.

[84] Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramón Fermandez, Silvio Amir, Luís Marujo, and Tiago Luís. Finding function in form: Compositional character models for open vocabulary word representation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1520–1530, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[85] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. Does gender matter? towards fairness in dialogue systems. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4403–4416, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[86] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. Does gender matter? towards fairness in dialogue systems, 2020.

[87] Haochen Liu, Tyler Derr, Zitao Liu, and Jiliang Tang. Say what I want: Towards the dark side of neural dialogue models. CoRR, abs/1909.06044, 2019.

[88] Haochen Liu, Wei Jin, Hamid Karimi, Zitao Liu, and Jiliang Tang. The authors matter: Understanding and mitigating implicit bias in deep text classification. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 74–85, Online, August 2021. Association for Computational Linguistics.

[89] Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. Mitigating gender bias for neural dialogue generation with adversarial learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 893–903, Online, November 2020. Association for Computational Linguistics.

[90] Edward Loper and Steven Bird. NLTK: The natural language toolkit. In Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, pages 63–70, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[91] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. CoRR, abs/1807.11714, 2018.

[92] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. In Logic, Language, and Security, pages 189–202. Springer, 2020.

[93] Juan M Madera, Michelle R Hebl, and Randi C Martin. Gender and letters of recommendation for academia: agentic and communal differences. Journal of Applied Psychology, 94(6):1591,

2009.

[94] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. CoRR, abs/1903.10561, 2019.

[95] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 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, Volume 1 (Long and Short Papers), pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[96] Paola Medel and Vahab Pournaghshband. Eliminating gender bias in computer science education materials. In Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education, pages 411–416, 2017.

[97] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. CoRR, abs/1908.09635, 2019.

[98] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635, 2019.

[99] Michela Menegatti and Monica Rubini. Gender bias and sexism in language. In Oxford Research Encyclopedia of Communication. 2017.

[100] Alexander H. Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. Parlai: A dialog research software platform. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017 - System Demonstrations, pages 79–84, 2017.

[101] Rishabh Misra. News category dataset, 06 2018.

[102] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. PLOS ONE, 15:1–26, 08 2020.

[103] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. arXiv preprint arXiv:1808.07231, 2018.

[104] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, 2014.

[105] Alexandra Guedes Pinto, Henrique Lopes Cardoso, Isabel Margarida Duarte, Catarina Vaz Warrot, and Rui Sousa-Silva. Biased language detection in court decisions. In International Conference on Intelligent Data Engineering and Automated Learning, pages 402–410. Springer, 2020.

[106] Barbara Plank, Anders Søgaard, and Yoav Goldberg. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 412–418, Berlin, Germany, August 2016. Association for Computational Linguistics.

[107] Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. Assessing gender bias in machine translation - A case study with google translate. CoRR, abs/1809.02208, 2018.

[108] Organizers of QueerInAI, Ashwin S, William Agnew, Hetvi Jethwani, and Arjun Subramonian. Rebuilding trust: Queer in ai approach to artificial intelligence risk management, 2021.

[109] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. CoRR, abs/1910.10683, 2019.

[110] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 856–865, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[111] Jan Riebling. Centering resonance analysis using nltk and networkx. http://www.sociology-hacks.org/?p=151, 2015.

[112] Alan Ritter, Colin Cherry, and William B. Dolan. Data-driven response generation in social media. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 583–593, 2011.

[113] James A Rodger and Parag C Pendharkar. A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. International Journal of Human-Computer Studies, 60(5-6):529–544, 2004.

[114] Adam Rose. Are face-detection cameras racist? Time Business, 2010.

[115] Michael J. Rosenfeld and Reuben J. Thomas. Searching for a mate: The rise of the internet as a social intermediary. American Sociological Review, 77(4):523–547, 2012.

[116] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. arXiv preprint arXiv:1804.09301, 2018.

[117] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14, New Orleans, Louisiana, June 2018. Association for

Computational Linguistics.

[118] Amrita Saha, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 705–713, 2018.

[119] Punyajoy Saha, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. "short is the road that leads from fear to hate": Fear speech in indian whatsapp groups. In Proceedings of the Web Conference 2021, WWW '21, page 1110–1121, New York, NY, USA, 2021. Association for Computing Machinery.

[120] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics.

[121] Roger C Schank. Conceptual dependency: A theory of natural language understanding. Cognitive psychology, 3(4):552–631, 1972.

[122] Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017.

[123] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the 30th AAAI Conference on Artificial Intelligence, pages 3776–3784, 2016.

[124] Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. Predictive biases in natural language processing models: A conceptual framework and overview. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5248–5264, Online, July 2020. Association for Computational Linguistics.

[125] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers, pages 1577–1586, 2015.

[126] Shanya Sharma, Manan Dey, and Koustuv Sinha. Evaluating gender bias in natural language

inference. CoRR, abs/2105.05541, 2021.

[127] Ian Stewart. Now we stronger than ever: African-American English syntax in Twitter. In Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 31–37, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[128] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112, 2014.

[129] Spencer S. Swinton. Predictive bias in graduate admissions tests. ETS Research Report Series, 1981, 1981.

[130] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks.

[131] Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2920–2935, Online, July 2020. Association for Computational Linguistics.

[132] Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. Mind your inflections! Improving NLP for non-standard Englishes with Base-Inflection Encoding. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5647–5663, Online, November 2020. Association for Computational Linguistics.

[133] Rachael Tatman and Conner Kasten. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In Proc. Interspeech 2017, pages 934–938, 2017.

[134] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019., pages 83–92, 2019.

[135] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 6000–6010, 2017.

[136] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics.

[137] David Manning White. The "gate keeper": A case study in the selection of news. Journalism

Quarterly, 1950.

[138] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122. Association for Computational Linguistics, 2018.

[139] Alden Williams. Unbiased study of television news bias. Journal of Communication, 1975.

[140] Marty J. Wolf, Keith W. Miller, and Frances S. Grodzinsky. Why we should have seen that coming: comments on microsoft's tay "experiment, " and wider implications. SIGCAS Computers and Society, 47(3):54–64, 2017.

[141] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. MIND: A large-scale dataset for news recommendation. In ACL, 2020.

[142] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. Demoting racial bias in hate speech detection. In Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, pages 7–14, Online, July 2020. Association for Computational Linguistics.

[143] Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. Detoxifying language models risks marginalizing minority voices. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2390–2397, Online, June 2021. Association for Computational Linguistics.

[144] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. CoRR, abs/1906.08237, 2019.

[145] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In Advances in Neural Information Processing Systems, pages 2921–2930, 2017.

[146] Wenjie Yin and Arkaitz Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science, 7:e598, 2021.

[147] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2:67–78, 2014.

[148] Vershawn Ashanti Young. "nah, we straight": An argument against code switching. JAC, 29(1/2):49–76, 2009.

[149] Vershawn Ashanti Young and Rusty Barrett. Other people's English: Code-meshing, code-switching, and African American literacy. Parlor Press LLC, 2018.

[150] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2015.

[151] Marcos Zampieri, Preslav Nakov, and Yves Scherrer. Natural language processing for similar languages, varieties, and dialects: A survey. Natural Language Engineering, 26(6):595–612, 2020.

[152] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. arXiv preprint arXiv:2004.14088, 2020.

[153] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4134–4145, Online, July 2020. Association for Computational Linguistics.

[154] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In NIPS, 2015.

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

[156] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. CoRR, abs/1804.06876, 2018.

[157] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods, 2018.

[158] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 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, pages 4847–4853, 2018.

[159] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. arXiv preprint arXiv:1809.01496, 2018.

[160] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. Examining gender bias in languages with grammatical gender, 2019.

[161] Xiang Zhou and Mohit Bansal. Towards robustifying NLI models against lexical dataset

biases. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8759–8771, Online, July 2020. Association for Computational Linguistics.

[162] Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. Challenges in automated debiasing for toxic language detection. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3143–3155, Online, April 2021. Association for Computational Linguistics.

[163] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics.

APPENDIX A

BIAS DETECTION IN DIALOGUE GENERATION

In the appendix, we detail the 6 categories of words, i.e., *gender (male and female), race (white and black), pleasant and unpleasant, career and family*.

A.1 Gender Words

The gender words consist of gender specific words that entail both male and female possessive words as follows:

(gods - goddesses), (nephew - niece), (baron - baroness), (father - mother), (dukes - duchesses), (dad - mom), (beau - belle), (beaus - belles), (daddies - mummies), (policeman - policewoman), (grandfather - grandmother), (landlord - landlady), (landlords - landladies), (monks - nuns), (stepson - stepdaughter), (milkmen - milkmaids), (chairmen - chairwomen), (stewards - stewardesses), (men - women), (masseurs - masseuses), (son-in-law - daughter-in-law), (priests - priestesses), (steward - stewardess), (emperor - empress), (son - daughter), (kings - queens), (proprietor - proprietress), (grooms - brides), (gentleman - lady), (king - queen), (governor - matron), (waiters - waitresses), (daddy - mummy), (emperors - empresses), (sir - madam), (wizards - witches), (sorcerer - sorceress), (lad - lass), (milkman - milkmaid), (grandson - granddaughter), (congressmen - congresswomen), (dads - moms), (manager - manageress), (prince - princess), (stepfathers - stepmothers), (stepsons - stepdaughters), (boyfriend - girlfriend), (shepherd - shepherdess), (males - females), (grandfathers - grandmothers), (step-son - step-daughter), (nephews - nieces), (priest - priestess), (husband - wife), (fathers - mothers), (usher - usherette), (postman - postwoman), (stags - hinds), (husbands - wives), (murderer - murderess), (host - hostess), (boy - girl), (waiter - waitress), (bachelor - spinster), (businessmen - businesswomen), (duke - duchess), (sirs - madams), (papas - mamas), (monk - nun), (heir - heiress), (uncle - aunt), (princes - princesses), (fiance - fiancee), (mr - mrs), (lords - ladies), (father-in-law - mother-in-law), (actor - actress), (actors - actresses), (postmaster - postmistress), (headmaster - headmistress), (heroes - heroines), (groom - bride), (businessman - businesswoman), (barons - baronesses), (boars - sows), (wizard - witch), (sons-in-law - daughters-in-law), (fiances

- fiancees), (uncles - aunts), (hunter - huntress), (lads - lasses), (masters - mistresses), (brother - sister), (hosts - hostesses), (poet - poetess), (masseur - masseuse), (hero - heroine), (god - goddess), (grandpa - grandma), (grandpas - grandmas), (manservant - maidservant), (heirs - heiresses), (male - female), (tutors - governesses), (millionaire - millionairess), (congressman - congresswoman), (sire - dam), (widower - widow), (grandsons - granddaughters), (headmasters - headmistresses), (boys - girls), (he - she), (policemen - policewomen), (step-father - step-mother), (stepfather - stepmother), (widowers - widows), (abbot - abbess), (mr. - mrs.), (chairman - chairwoman), (brothers - sisters), (papa - mama), (man - woman), (sons - daughters), (boyfriends - girlfriends), (he's - she's), (his - her).

A.2 Race Words

The race words consist of Standard US English words and African American/Black words as follows: (going - goin), (relax - chill), (relaxing - chillin), (cold - brick), (not okay - tripping), (not okay - spazzin), (not okay - buggin), (hang out - pop out), (house - crib), (it's cool - its lit), (cool - lit), (what's up - wazzup), (what's up - wats up), (what's up - wats popping), (hello - yo), (police - 5-0), (alright - aight), (alright - aii), (fifty - fitty), (sneakers - kicks), (shoes - kicks), (friend - homie), (friends - homies), (a lot - hella), (a lot - mad), (a lot - dumb), (friend - mo), (no - nah), (no - nah fam), (yes - yessir), (yes - yup), (goodbye - peace), (do you want to fight - square up), (fight me - square up), (police - po po), (girlfriend - shawty), (i am sorry - my bad), (sorry - my fault), (mad - tight), (hello - yeerr), (hello - yuurr), (want to - finna), (going to - bout to), (That's it - word), (young person - young blood), (family - blood), (I'm good - I'm straight), (player - playa), (you joke a lot - you playing), (you keep - you stay), (i am going to - fin to), (turn on - cut on), (this - dis), (yes - yasss), (rich - balling), (showing off - flexin), (impressive - hittin), (very good - hittin), (seriously - no cap), (money - chips), (the - da), (turn off - dub), (police - feds), (skills - flow), (for sure - fosho), (teeth - grill), (selfish - grimey), (cool - sick), (cool - ill), (jewelry - ice), (buy - cop), (goodbye - I'm out), (I am leaving - Imma head out), (sure enough - sho nuff), (nice outfit - swag), (sneakers - sneaks), (girlfiend - shortie), (Timbalands - tims), (crazy - wildin), (not cool - wack), (car - whip), (how are you - sup), (good - dope), (good - fly), (very good - supafly), (prison - pen), (friends -

squad), (bye - bye felicia), (subliminal - shade).

A.3 Pleasant and Unpleasant Words

Pleasant words. The pleasant words consist of words often used to express positive emotions and scenarios as follows:

caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation, joy, wonderful.

Unpleasant Words. The unpleasant words consist of words often used to express negative emotions and scenarios as follows:

abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison, terrible, horrible, nasty, evil, war, awful, failure.

A.4 Career and Family Words

Career Words. The career words consist of words pertain to careers, jobs and businesses:

company, industry, academic, executive, management, occupation, professional, corporation, salary, office, business, career, technician, accountant, supervisor, engineer, worker, educator, clerk, counselor, inspector, mechanic, manager, therapist, administrator, salesperson, receptionist, librarian, advisor, pharmacist, janitor, psychologist, physician, carpenter, nurse, investigator, bartender, specialist, electrician, officer, pathologist, lawyer, planner, practitioner, plumber, instructor, surgeon, veterinarian paramedic, examiner, chemist, machinist, appraiser, nutritionist, architect, hairdresser, baker, programmer, paralegal, hygienist, scientist.

Family Words. The family words consist of words refer to relations within a family or group of people.

adoption, adoptive, birth, bride, bridegroom, care-giver, child, childhood, children, clan, cousin, devoted, divorce, engaged, engagement, estranged, faithful, family, fiancee, folks, foster, groom, heir, heiress, helpmate, heritage, household, husband, in-law, infancy, infant, inherit, inheritance,

kin, kindred, kinfolk, kinship, kith, lineage, love, marry, marriage, mate, maternal, matrimony, natal, newlywed, nuptial, offspring, orphan, parent relative, separation, sibling, spouse, tribe, triplets, twins, wed, wedding, wedlock, wife.

APPENDIX B

DETECTING AND EXAMINING GENDER BIAS IN THE NEWS

B.1 Appendix A

As previously mentioned, we provide one of the largest non offensive, non repeating set of gender-specific (*male* and *female*) words, we will now detail the 2 categories containing a total of 465 masculine and feminine gender possessive nouns. Note that in the creation of the set of words, overly offensive gender related words such as *bitch, whore, slut, bastard, prick, etc.,* were left out of the sets of nouns as they are hardly ever used in news articles. However, offensive gender related words are often used in tabloids (a compact version of a newspapers dominated by headline titles and images). [69].

B.1.1 Male Possessive Words

The succeeding word list consists of 230 gender specific words that entail *male* possessive nouns as follows:

god, gods, nephew, nephews, baron, father, fathers dukes, dad, beau, beaus, daddies, policeman, policemen, grandfather, landlord, landlords, monk, monks, step-son, step-sons, milkmen, chairmen, chairman, steward, men, masseurs, son-in-law, priest, king, governor, waiter, daddy, steward, emperor, son, proprietor, groom, grooms, gentleman, gentlemen, sir, wizards, sorcerer, lad, milk-man, grandson, grand-son, congressmen, dads, manager, prince, stepfathers, boyfriend, shepherd, shepherds, males, grandfathers, grand-fathers, husband, usher, postman, stags, husbands, host, boy, waiter, bachelor, bachelors, businessmen, duke, sirs, papas, heir, uncle, princes, fiance, mr, lords, father-in-law, actor, actors, postmaster, headmaster, heroes, businessman, boars, wizard, sons-in-law, fiances, uncles, hunter, lads, masters, brother, hosts, poet, hero, grandpa, grandpas, manservant, heirs, male, tutors, millionaire, congressman, sire, sires, widower, grandsons, grand-sons, boys, he, step-father, jew, bridegroom, bridegrooms stepfather, widowers, abbot, mr., brothers, man, sons, boyfriends, he's, his, him, earl, giant, count, stepson, stepsons, poet, mayor, peer, negro, abbot, traitor, benefactor, instructor, conductor, founder, founders, hunters, huntresses, temptress,

enchanter, enchanters, songster, songsters, murderer, murderers, patron, patrons, author, czar, guy, spokesman, spokesmen, pa, councilman, council-man, councilmen, council-men, gay, gays, prostate cancer, fraternity, fraternities, salesman, dude, dudes, paternal, brotherhood, statesman, statesmen, countryman, countrymen, suitor, macho, papa, strongman, strongmen, boyhood, manhood, masculine, macho, horsemen, brethren, chap, chaps, schoolboy, schoolboys, bloke, blokes, patriarch, patriachy, fatherhood, hubby, hubbies, fella, fellas, handyman, fraternal, bro, masculinity, ballerino, pappy, papi, pappies, dada, bf, bfs, knights, knight, menfolk, brotherly, manly, pimp, pimps, homeboy, homeboys, grandnephew, grand-nephew, grand-nephew, grand-nephews, john doe, nobleman, noblemen, dream boy, himself, gramps

B.1.2 Female Possessive Words

The succeeding word list consists of 235 gender specific words that entail *female* possessive nouns as follows:

goddesses, niece, baroness, mother, duchesses, mom, belle, belles, mummies, policewoman, grandmother, landlady, landladies, nuns, stepdaughter, milkmaids, chairwomen, stewardesses, women, masseuses, daughter-in-law, priestesses, stewardess, empress, daughter, queens, proprietress, brides, lady, queen, matron, waitresses, mummy, empresses, madam, witches, sorceress, lass, milkmaid, granddaughter, grand-daughter, congresswomen, moms, manageress, princess, stepmothers, stepdaughters, girlfriend, shepherdess, females, grand-mothers, grandmothers, step-daughter, nieces, priestess, wife, mother, usherette, postwoman, hind, wives, murderess, hostess, girl, waitress, spinster, shepherdess, businesswomen, duchess, madams, mamas, nun, heiress, aunt, princesses, fiancee, mrs, ladies, mother-in-law, actress, actresses, postmistress, headmistress, heroines, bride, businesswoman, baronesses, sows, witch, daughters-in-law, aunts, huntress, lasses, mistress, mistresses, sister, hostesses, poetess, masseuse, heroine, goddess, grandma, grandmas, maidservant, heiresses, patroness, female, governesses, millionairess, congresswoman, dam, widow, granddaughters, grand-daughters, headmistresses, girls, she, policewomen, step-mother, stepmother, widows, abbess, mrs., chairwoman, sisters, mama, woman, daughters, girlfriends, she's, her, maid, countess, giantess, poetess, jewess, mayoress, peeress, negress, abbess, traitress, benefactress, instructress,

conductress, founder, huntress, temptress, enchantress, songstress, murderess, murderesses, patronesses, authoress, czarina, spokeswoman, spokeswomen, ma, councilwoman, council-woman, councilwomen, council-women, mum, lesbian, lesbians, breast, breasts, maiden, maidens, sorority, sororities, saleswoman, dudette, maternal, feminist, feminists, sisterhood, housewife, housewives, stateswoman, stateswomen, countrywoman, countrywomen, chick, chicks, mommy, strongwoman, strongwomen, babe, babes, diva, divas, feminine, feminism, gal, gals, sistren, schoolgirl, schoolgirls, matriarch, matriarchy, motherhood, wifey, sis, femininity, ballerina, ballerinas, granny, grannies, mami, momma, maam, gf, gfs, damsel, damsels, vixen, vixens, nan, nanny, nannies, auntie, womenfolk, sisterly, motherly, homegirl, homegirls, grand-niece, grand-nieces, grandniece, grandnieces, jane doe, noblewoman, noblewomen, dream girl, madame, herself, hers

B.2 Appendix B

As previously mentioned, we provide one of the largest gender-specific and gender-neutral words containing a total of 357 masculine, feminine and neutral career-related and family-related words, we will now we will now detail the 2 categories of family-related and career-related words.

B.2.1 Career Words

The succeeding word list consists of 162 gender specific and gender neutral career-related words as follows:

policewoman, milkmaids, chairwomen, stewardesses, masseuses, priestesses, stewardess, proprietress, waitresses, congresswomen,moms, manageress, shepherdess, priestess, usherette, postwoman, hostess, waitress, spinster, shepherdess, businesswomen, actress, actresses, postmistress, headmistress, huntress, mistress, mistresses, sister, hostesses, masseuse,maidservant, heiresses, patroness, governesses, congresswoman, headmistresses, policewomen, chairwoman, maid, mayoress, peeress, traitress, benefactress, instructress, conductress, huntress,

temptress, enchantress, songstress, spokeswoman, spokeswomen, councilwoman, council-woman, councilwomen, council-women,

saleswoman, stateswoman, stateswomen, policeman, policemen, landlord, landlords, chairmen,

chairman, steward, priest, king, governor, waiter, steward, proprietor, sorcerer, congressmen, dads, manager, waiter, actor, actors, postmaster, headmaster, businessman, manservant, tutors, congressman, benefactor, instructor, conductor, founder, founders, hunters, huntresses, tempt, enchanter, enchanters, spokesman, spokesmen, councilman, council-man, councilmen, council-men, salesman, handyman, knights, knight, academic, accountant, administrator, advisor, appraiser, architect, baker, bartender, business, career, carpenter, chemist, clerk, company, corporation, counselor, educator, electrician, engineer, examiner, executive, hairdresser, hygienist, industry, inspector, instructor, investigator, janitor, lawyer, librarian, machinist, management, mechanic, nurse, nutritionist, occupation, officer, paralegal, paramedic, pathologist, pharmacist, physician, plumber, practitioner, programmer, psychologist, receptionist, salary, salesperson, scientist, specialist, supervisor, surgeon, technician, therapist, veterinarian, worker

B.2.2 Family Words

The succeeding word list consists of 195 gender specific and gender neutral family-related words as follows:

niece, mother, mom, mummies, grandmother, nuns, stepdaughter, women, daughter-in-law, daughter, queens, brides, mummy, empresses, madam, granddaughter, grand-daughter, moms, stepmothers, stepdaughters, girlfriend, grand-mothers, grandmothers, step-daughter, nieces, wife, mothers, wives, girl, madams, mamas, aunt, fiancee, mrs, mother-in-law, bride, daughters-in-law, aunts, heir, heiress, sister, grandma, grandmas, dam, widow, granddaughters, grand-daughters, girls, she, step-mother, stepmother, mrs., sisters, mama, woman, daughters, girlfriends, ma, mum, mommy, gal, gals, sistren, matriarch, matriarchy, motherhood, wifey, sis, granny, grannies, mami, momma, ma'am, gf, gfs, damsel, damsels, vixen, vixens, nanny, nannies,auntie, womenfolk, sisterly, motherly, homegirl, homegirls, grand-niece, grand-nieces, grandniece, grandnieces, madame, him, father, fathers, dad, beau, beaus, daddies, grandfather, step-son, step-sons, men, son-in-law, daddy, son, groom, grooms, sir, grandson, grand-son, dads, prince, stepfathers,boyfriend, grandfathers, grand-fathers, husband, husbands, boy, bachelor, bachelors, sirs, papas, uncle, princes, fiance, mr, father-in-law, sons-in-law, fiances, uncles, brother, grandpa, grandpas, widower, grandsons,

grand-sons,boys, step-father, bridegroom, bridegrooms, stepfather, widowers, mr., brothers, man, sons, boyfriends, he's, his, stepson, stepsons, guy, fraternity, fraternities, salesman, dude, dudes, paternal, brotherhood, papa, boyhood, manhood, masculine, brethren, chap, chaps, patriarch, patriachy, fatherhood, hubby, hubbies, fella, fellas, fraternal, bro, pappy, papi, pappies, dada, bf, bfs, brotherly, homeboy, homeboys,grandnephew, grand-nephew, grand-nephew, grand-nephews, gramps, family, infancy, infant, kin, orphan, twin

APPENDIX C

DETECTING HARMFUL ONLINE CONVERSATIONAL CONTENT TOWARDS LGBTQIA2S+ INDIVIDUALS

C.1 Annotation Guidelines

First, you will be given an extensive list of acronyms and terms from OutRight (Link: https://outrightinternational.org/c explained), an LGBTQIA2S+ human rights organization. After, you will be given a comment, where your task is to indicate whether a comment is *toxic* or *non-toxic*. If a comment is deemed *toxic*, then you be provided with 5 additional labels (*severe toxicity*, *obscene*, *threat*, *insult* and *identity attack*) to correctly identify and determine if a comment qualifies to be classified under one or more of the 5 additional labels.

Human Annotator Protocol

1. Are you a member of the LGBTQIA2S+ community?
2. If you responded "no" above, are you an LGBTQIA2S+ activist or ally?
3. If you responded "no" above, please stop here.
4. If you responded "yes" any of above question, given the extensive acronym list what is your identity, sexuality, and relationship? (Optional. This information is collected, but not saved, only for demographic purposes.)
5. Are you willing to annotate several Reddit comments that contain *stereotypes*, *profanity*, *vulgarity* and other harmful language geared towards LGBTQIA2S+ individuals?
6. If you responded "yes" above, we must mention that if you believe you may become triggered or disturbed and cannot continue, please stop here.
7. If you responded "no" above, please stop here.

Rating/ Sensitivity Protocol

1. As you responded "yes" a previous question,

... Are you willing to annotate several Reddit comments that contain stereotypes, profanity, vulgarity and other harmful language geared towards LGBTQIA2S+ individuals?

You will be provided with 1000 comments which we have sampled from our *binary* classification, and 5 additional labels.

2. For each comment, you will be tasked is to indicate whether a comment is *toxic* or *non-toxic*. Is this comment toxic?

    a) If you responded "yes" above, please select one or more of the appropriate labels provided considering these two classes, "harmful : 1" and "non-harmful : 0".

    b) If you responded "no" above, please discard this comment.

3. Have you ever seen, heard, used or been called any of these Anti-LGBTQIA2S+ terms in a particular comment, for example, on social media or in-person?

4. If you responded "yes" above, do you feel triggered, disturbed or distressed reading this comment. (Optional. This information is collected, but is not saved, only for demographic purposes.)

We would like to remind you that the objective of this study is not to cause more harm, but to create a safe and inclusive place that welcomes, supports, and values all LGBTQIA2S+ individuals both online and offline. However, due to the overall purpose of this study, we focus on online inclusivity.

C.2 Data Breakdown

In this section, we display a breakdown of the data as the toxicity label is not an across-the-board label, but there exists a large amount of overlap between labels.

In total, there are 7459 toxicity comments. (75.12% of all data.)
- 185 or 2.48% were also severe toxicity.
- 1590 or 21.32% were also obscene.

- 28 or 0.38% were also threat.
- 2244 or 30.08% were also insult.
- 4494 or 60.25% were also identity attack.

In total, there are 185 severe toxicity comments. (1.86% of all data.)
- 185 or 100.00% were also toxicity.
- 185 or 100.00% were also obscene.
- 13 or 7.03% were also threat.
- 185 or 100.00% were also insult.
- 184 or 99.46% were also identity attack.

In total, there are 1590 obscene comments. (16.01% of all data.)
- 1590 or 100.00% were also toxicity.
- 185 or 11.64% were also severe toxicity.
- 23 or 1.45% were also threat.
- 1512 or 95.09% were also insult.
- 1443 or 90.75% were also identity attack.

In total, there are 28 threat comments. (0.28% of all data.)
- 28 or 100.00% were also toxicity.
- 13 or 46.43% were also severe toxicity.
- 23 or 82.14% were also obscene.
- 25 or 89.29% were also insult.
- 27 or 96.43% were also identity attack.

In total, there are 2244 insult comments. (22.60% of all data.)
- 2244 or 100.00% were also toxicity.

- 185 or 8.24% were also severe toxicity.

- 1512 or 67.38% were also obscene.

- 25 or 1.11% were also threat.

- 2141 or 95.41% were also identity attack.

In total, there are 4494 identity attack comments. (45.26% of all data.)

- 4494 or 100.00% were also toxicity.

- 184 or 4.09% were also severe toxicity.

- 1443 or 32.11% were also obscene.

- 27 or 0.60% were also threat.

- 2141 or 47.64% were also insult.

C.3 Feature Distribution Plots

In this section, we display feature distribution *i.e.*, visualization of the variation in the data distribution of each label. These distribution plots represent the overall distribution of the continuous data variables.

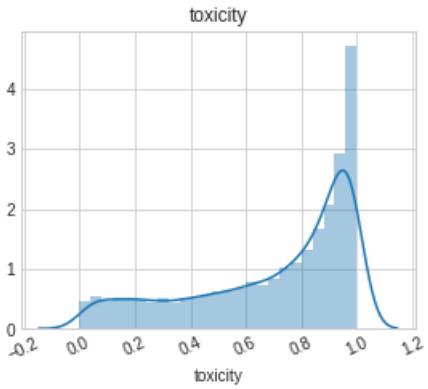

Figure C.1: Toxicity feature distribution.

C.4 Word Contribution

Disclaimer: Due to the overall purpose of the study, several terms in the figures may be offensive or disturbing (e.g. profane, vulgar, or homophobic slurs). These terms are not filtered as they are

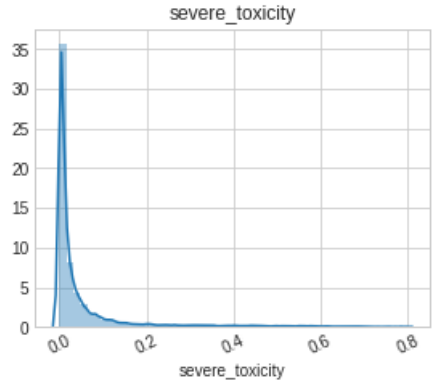

Figure C.2: Severe Toxicity feature distribution.

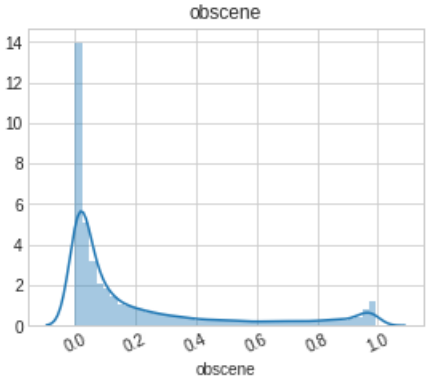

Figure C.3: Obscene feature distribution.

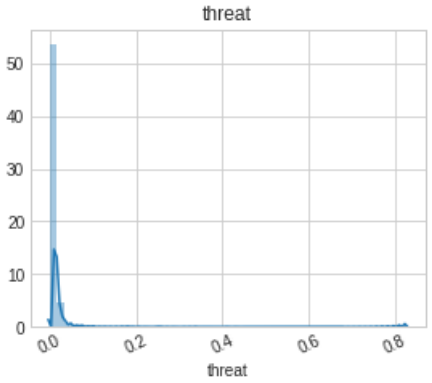

Figure C.4: Threat feature distribution.

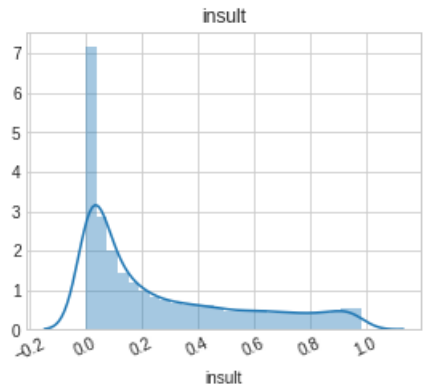

Figure C.5: Insult feature distribution.

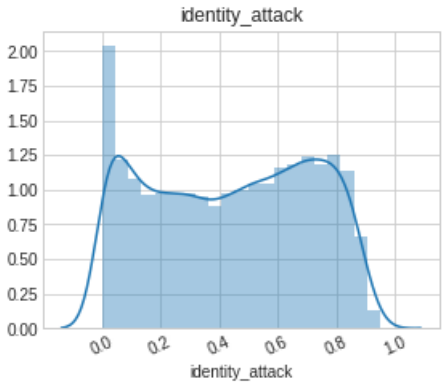

Figure C.6: Identity attack feature distribution.

representative of essential aspects in the dataset.

In this section, we demonstrate which words constitutes towards a “harmful” or “non-harmful” comment. In Figures C.7 – C.12, we display the top 30 most frequent words per label.

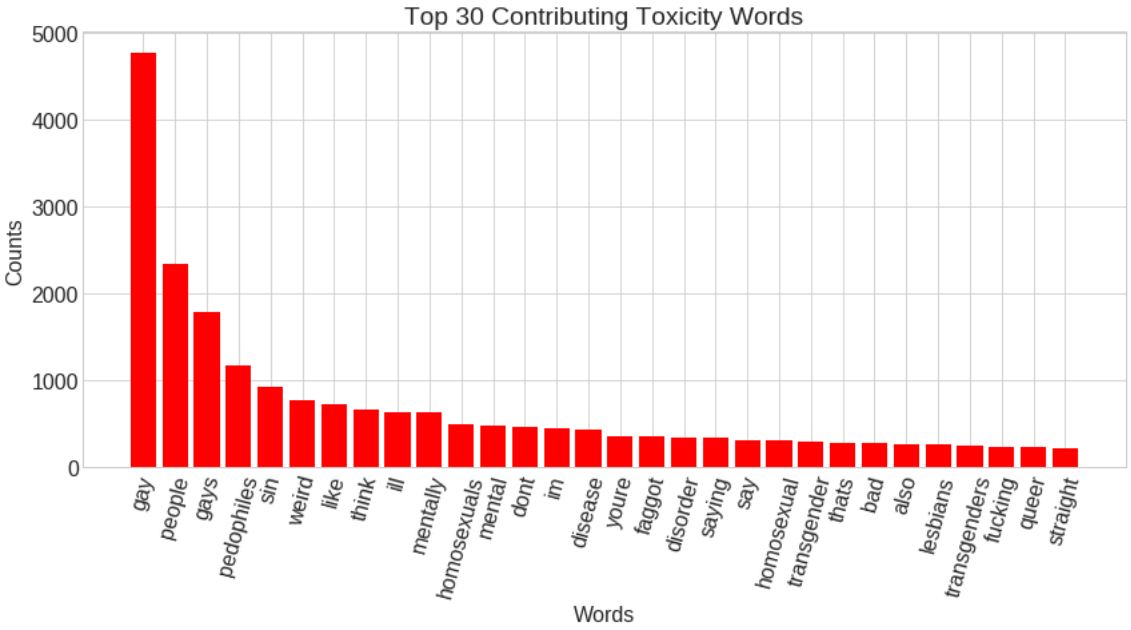

Figure C.7: Top 30 most frequent words contributing to the Toxicity label.

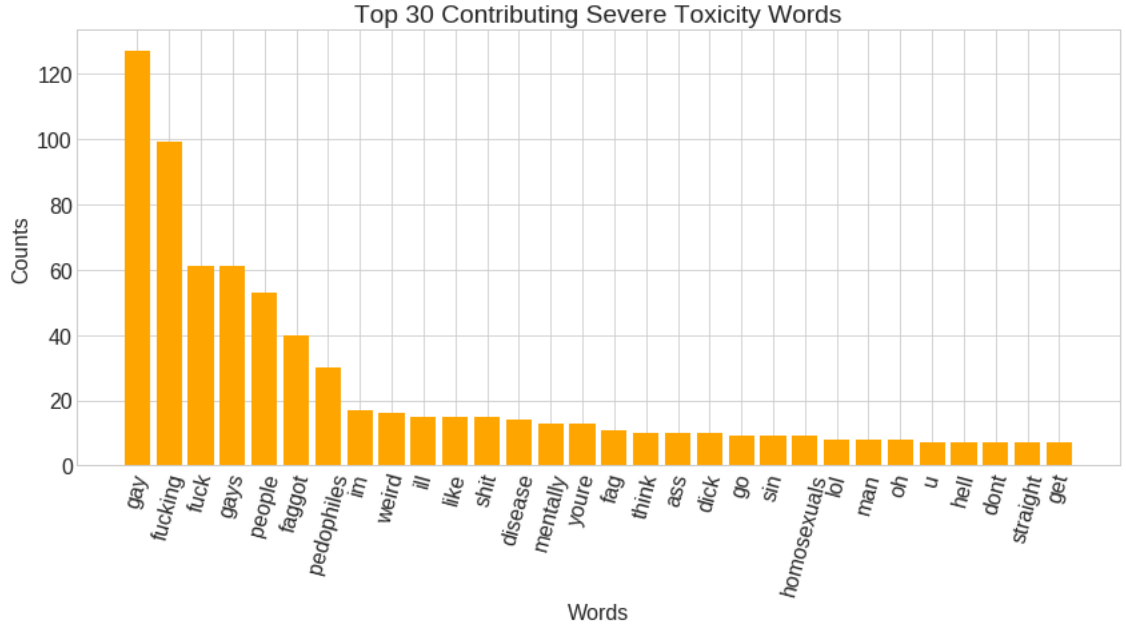

Figure C.8: Top 30 most frequent words contributing to the Severe Toxicity label.

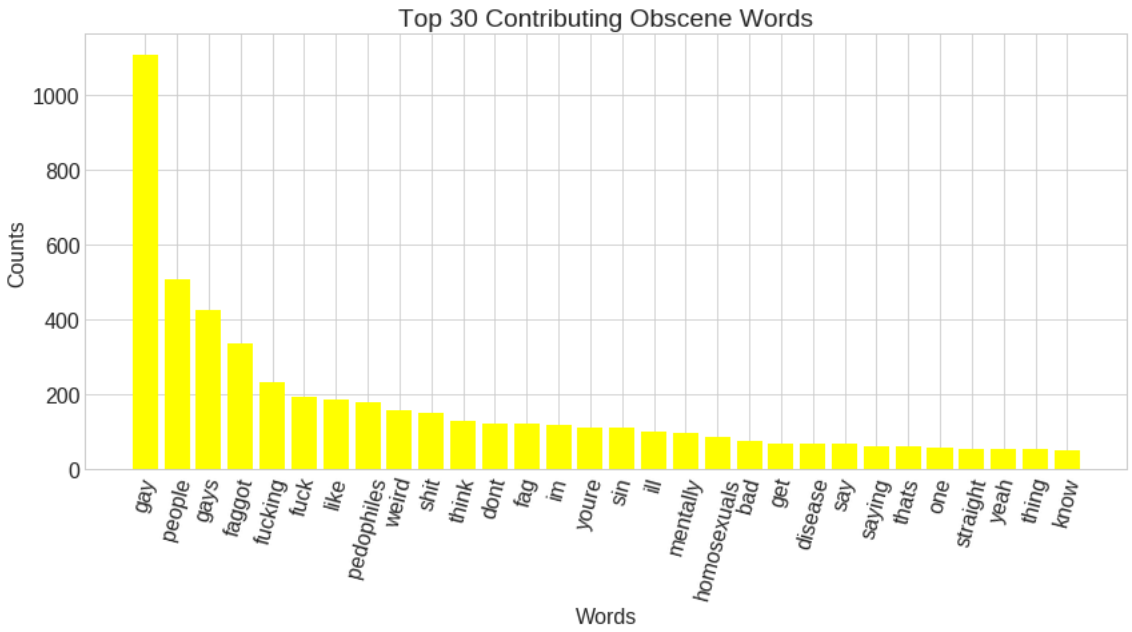

Figure C.9: Top 30 most frequent words contributing to the Obscene label.

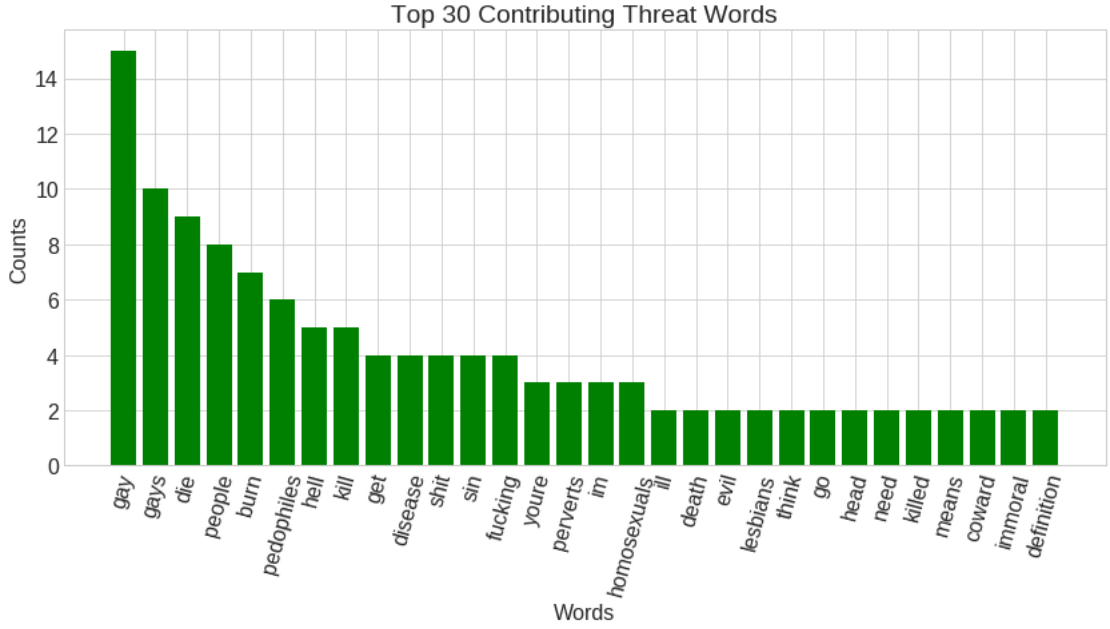

Figure C.10: Top 30 most frequent words contributing to the Threat label.

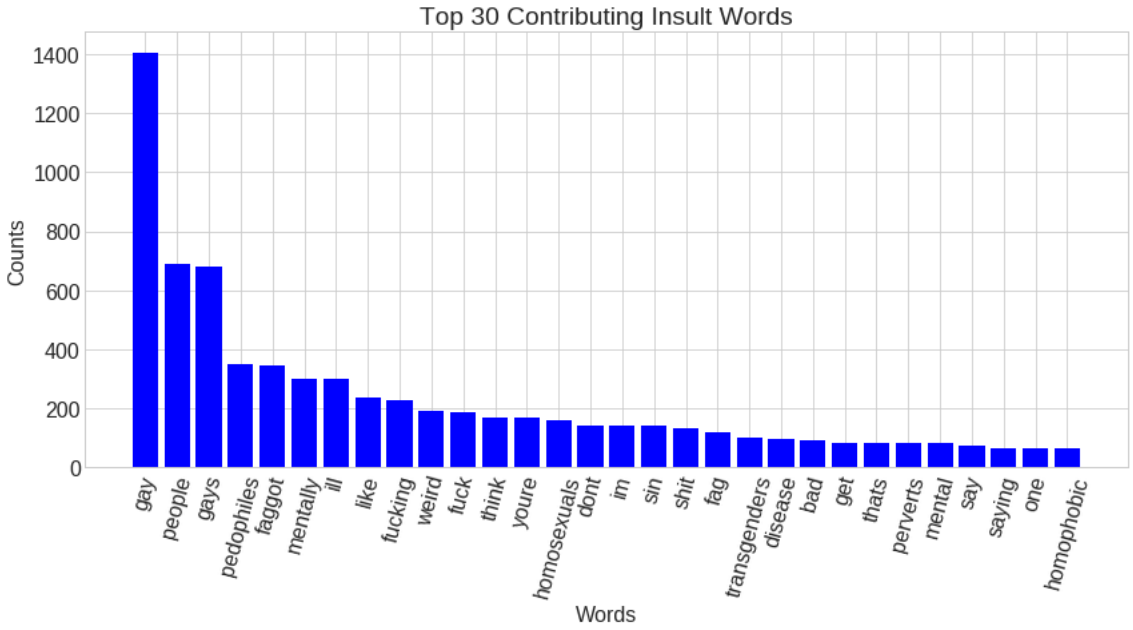

Figure C.11: Top 30 most frequent words contributing to the Insult label.

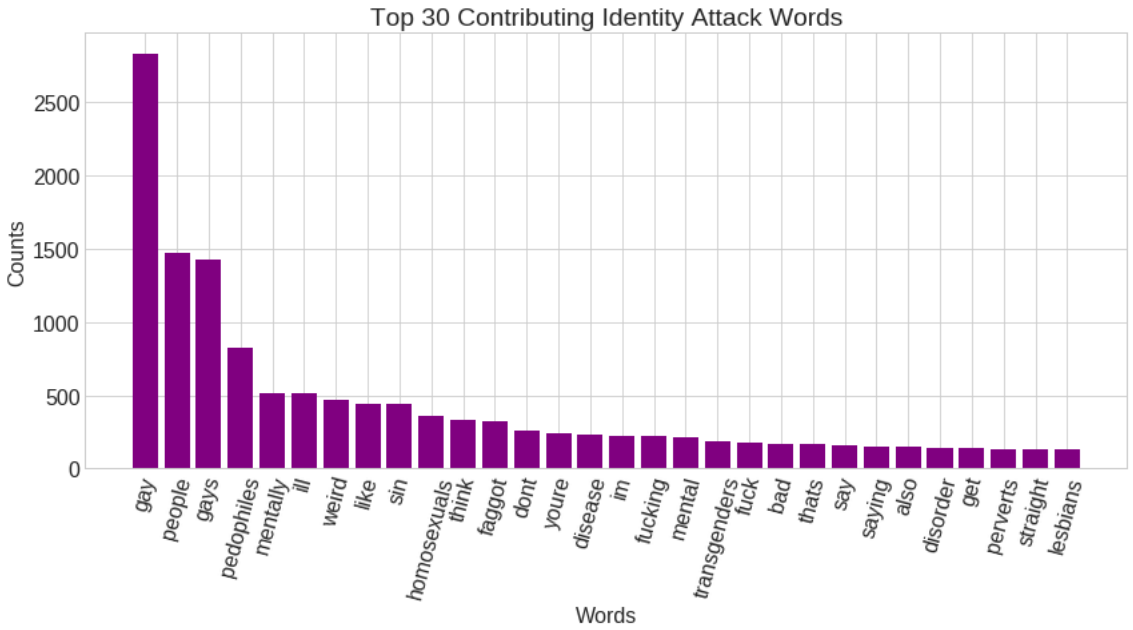

Figure C.12: Top 30 most frequent words contributing to the Identity Attack label.

APPENDIX D

A MULTI-LAYERED LANGUAGE ANALYSIS: A CASE STUDY OF AFRICAN-AMERICAN ENGLISH

D.1 Dataset Details

Our collected dataset is demographically-aligned on AAE in correspondence on the dialectal tweet corpus by [14]. The TwitterAAE corpus is publicly available and can be downloaded from: http://slanglab.cs.umass.edu/TwitterAAE/. [14] uses a mixed-membership demographic language model which calculates demographic dialect proportions for a text accompanied by a race attribute—African America, Hispanic, Other, and White in that order. The race attribute is annotated by a jointly inferred probabilistic topic model based on the geolocation information of each user and tweet. Given that geolocation information (residence) is highly associated with the race of a user, the model can make accurate predictions. However, there a a low number messages that possess a posterior probabilities of NaN as these are messages that have no in-vocabulary words under the model.

D.2 Annotator Annotation Guidelines

You will be given demographically-aligned African American tweets, in which we refer to these tweets as sequences. As a dominant AAE speaker, who identifies as bi-dialectal, your task is to correctly identify the context of each word in a given sequence in hopes to address the issues of lexical, semantic and syntactic ambiguity.

1. Are you a dominant AAE speaker?

2. If you responded “yes” above, are you bi-dialectal?

3. If you responded “yes”, given a sequence, have you ever said, seen or used any of these words given the particular sequence?

4. Given a sequence, what are the SAE equivalents to the identified non-SAE terms?

5. For morphological and phonological (dialectal) purposes, are these particular words spelt how would you say or use them?

6. If you responded "no" above, can you provide a different spelling along with its SAE equivalent?

D.2.1 Annotation Protocol

1. What is the context of each word given the particular sequence?

2. Given NLTK's Penn Treebank Tagset, what is the most appropriate POS tag for each word in the given sequence?

D.2.2 Human evaluation of POS tags Protocol

1. Given the tagged sentence, are there any misclassified tags?

2. If you responded "yes" above, can you provide a different POS tag, and state why it is different?

D.3 Variable Rules Examples

In this section we present a few examples of simple, deterministic phonological and morphological language features or *current* variable rules which highlight several regional varieties of AAE which typically attain misclassified POS tags. Please note that a more exhaustive list of these rules is still being constructed as this work is still ongoing. Below are a few variable cases (MAE → AAE), some of which may have been previously shown in Table 5.2:

1. Consonant ('**t**') deletion (Adverb case) : e.g. "*just*" → "*jus*"; "*must*" → "*mus*"

2. Contractive negative auxiliary verbs replacement: "*doesn't*" → "*don't*"

3. Contractive (**'re**) loss: e.g. "*you're*" → "*you*"; "*we're*" → "*we*"

4. Copula deletion: Deletion of the verb "**be**" and its variants, namely "**is**" and "**are**" e.g. "*He* **is** *on his way*" → "*He on his way*"; "*You* **are** *right*" → "*You right*"

5. Homophonic word replacement (Pronoun case): e.g. "*you're*" → "*your*"

6. Indefinite pronoun replacement: e.g. "*anyone*" → "*anybody*";

7. Interdental fricative loss (Coordinating Conjuction case): e.g. "*this*" → "*dis*"; '*that*' → '*dat*"; "*the*" → "*da*"

8. Phrase reduction (present/ future tense) ⇒ word (Adverb case): e.g. "*what's up*" → "*wassup*"; "*fixing to*" → "*finna*"

9. Present tense possession replacement: e.g. "*John **has** two apples*" → "*John **got** two apples*"; "*The neighbors **have** a bigger pool*" → "*The neighbors **got** a bigger pool*"

10. Remote past "***been***" + completive ('**done**'): "*I've **already done** that*" → "*I **been done** that*"

11. Remote past "***been***" + completive ('**did**'): "*She **already did** that*" → "*She **been did** that*"

12. Remote past "***been***" + Present tense possession replacement: "*I **already have** food*" → "*I **been had** food*"; "*You **already have** those shoes*" → "*You **been got** those shoes*"

13. Term-fragment deletion: e.g. "*brother*" → "*bro*"; "*sister*" → "*sis*"; "*your*" → "*ur*"; "*suppose*" → "*pose*"; "*more*" → "*mo*"

14. Term-fragment replacement: "*something*" → "*sumn*"; "*through*" → "*thru*"; "*for*" → "*fa*"; "*nothing*" → "*nun*"

APPENDIX E

DETECTING AND MITIGATING INHERENT LINGUISTIC BIAS IN LARGE LANGUAGE MODELS

E.1 Implementation Details

E.1.1 Details of the Base Model

BERT – Bidirectional Encoder Representations from Transformers (BERT) [41] is a Transformer-based ML technique for NLP that achieves state-of-the-art results in a wide variety of NLP tasks. BERT is trained on a huge Books Corpus + Wikipedia dataset i.e., raw unlabeled English text consisting of 3.3 billion words. This model exploits an attention mechanism to learn contextual relationships between words and optimizes two objectives: (1) Masked Language Modeling (MLM) and (2) Next Sentence Prediction (NSP), and has a vocabulary size of 30,522. **Notation.** Given a sequence or sub-word tokens, for example, a sentence, $X = (x_1, x_2, \ldots, x_n)$, BERT trains an encoder which generates contextualized vector representations for each word-token: $Encoder(x_1, x_2, \ldots, x_n) = \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.

Masked Language Model. Also known as a *cloze* test, is the task of predicting missing tokens in a sequence when replaced with a [MASK] token. Specifically, to predict a subset of tokens $Y \subseteq X$ when sampled and substituted for a different tokens. Hence, the task is to predict the original tokens in Y from the altered input. Note that BERT selects each token in Y independently by randomly selecting a subset.

Next Sentence Prediction. The task of NSP is to jointly utilize two sequences (X_A, X_B) in a bi-sequence sampling procedure and predict whether X_B is a undeviating continuation of X_A. BERT first reads X_A, and then reads X_B in one of two ways: (1) reading X_B directly after X_A has ended; or (2) randomly sampling X_B from the corpus. To form X_A, X_B as an input to BERT, a [SEP] token is added to separate both sequence, and a special [CLS] token is added, where the target of [CLS] is to determine if X_B indeed follows X_A in the corpus.

E.1.2 Details of Experimental Settings

In summary, BERT optimizes its two objectives uniformly, and thus, it serves as a appropriate model for our task of understanding the inferential relationships between sentence pairs by examining the differences in language styles from different demographic groups e.g. African Americans. Now, we will now give details of each pretrained BERT model below:

1. BERT-base-uncased - Trained on raw English text, and consists of 12-layers, 768-hidden, 12-heads, 110M parameters.

2. BERT-large-cased - Trained on raw lower-cased English text, and consists of 24-layer, 1024-hidden, 16-heads, 335M parameters. Trained on cased English text.

E.2 Translative Morpho-syntax Protocol

Here we present a set of 20 linguistic phonetic and morphological text rules that are used to *code-switch* from SAE to AAE while maintaining contextual accuracy i.e., original structure, intent, semantic equivalence, and quality of a text. Please note that these are only a few examples of the most commonly used morphological linguistic AAE features (which we adapt from AAE literature). Our deterministic translative morpho-syntax protocol (TMsP) and its cases are as follows:

1. Consonant (‘**t**’) deletion (Special case) : e.g. “*just*” → “*jus*”; “*must*” → “*mus*”

2. Contractive (**’all**) gain: “*You all*” → “*Y***’all**”

3. Contractive negative auxiliary verbs replacement: “*doesn’t*” → “*don’t*”

4. Contractive (**’re**) loss: e.g. “*you’re*” → “*you*”; “*we’re*” → “*we*”; “*they’re*” → “*they*”

5. Contractive word replacement: e.g. “*isn’t*” → “*ain’t*”; “*wasn’t*” → “*ain’t*”

6. Copula deletion: Deletion of the verb “**be**” and its variants, namely “**is**” and “**are**” e.g. “*He* **is** *on his way*” → “*He on his way*”; “*You* **are** *right*” → “*You right*”

7. Gerund consonant (‘**g**’) deletion and retainment:

- Consonant (‘**g**’) deletion: e.g. “*coming*” → “*comin*”; “*going*” → “*goin*”
- Consonant (‘**g**’) retainment (Exception case): e.g. “*–ingin**g***”

8. Homophonic word replacement: e.g. “*whine*” → “*wine*”; “*you’re*” → “*your*”
9. Indefinite article replacement: e.g. “*an*” → “*a*”
10. Indefinite pronoun replacement: e.g. “*anyone*” → “*anybody*”; “*everyone*” → “*everybody*”
11. Interdental fricative loss: e.g. “*this*” → “*dis*”; ‘*that*’ → ‘*dat*”; “*than*” → “*dan*”; “*their*” → “*they* (*dey*)”; “*the*” → “*da*”
12. Negative concord replacement: e.g. “*Don’t say **anything***” → “*Don’t say **nothing***”
13. Phrase reduction (present/ future tense) ⇒ word e.g. “*going to*” → “*gonna*”; “*want to*” → “*wanna*”; “*trying to*” → “*tryna*”; “*what’s up*” → “*wassup*”; “*fixing to*” → “*finna*”
14. Possessive (**’s**) removal: e.g. “*He**’s** mad at me*” → “*He mad at me*”
15. Present tense possession replacement: e.g. “*John **has** two apples*” → “*John **got** two apples*”; “*The neighbors **have** a bigger pool*” → “*The neighbors **got** a bigger pool*”
16. Remote past “***been***” + completive (‘**done**’): “*I’ve **already done** that*” → “*I **been done** that*”
17. Remote past “***been***” + completive (‘**did**’): “*She **already did** that*” → “*She **been did** that*”
18. Remote past “***been***” + Present tense possession replacement: “*I **already have** food*” → “*I **been had** food*”; “*You **already have** those shoes*” → “*You **been got** those shoes*”
19. Term-fragment deletion: e.g. “*brother*” → “*bro*”; “*sister*” → “*sis*”; “*your*” → “*ur*”; “*suppose*” → “*pose*”; “*more*” → “*mo*”
20. Term-fragment replacement: “*something*” → “*sumn*”; “*through*” → “*thru*”; “*for*” → “*fa*”; “*nothing*” → “*nun*”

E.3 Annotation Guidelines

You will be given a phrase that is written in Standard American English (SAE), your task is to correctly identify if the translative vocabulary rules in Appendix E.2 are accurate in order to translate SAE text to AAE text. Furthermore, while reviewing the rules, be sure to mention that these rules and/or morpho-syntax word cases in the sampled premise-hypothesis sentence pairs maintain their contextual accuracy i.e., original structure, intent, semantic equivalence, and quality.

SAE to AAE Protocol

1. Are you a dominant AAE speaker?
2. If you responded "yes" above, are you bi-dialectal?
3. If you responded "yes" above, are you capable of code-switching by alternating between SAE and AAE frequently on a daily basis in a single conversation or situation?
4. Given TMsP above in Appendix E.2, are these main grammatical, structural and syntactic rules of word case usage of AAE linguistic features?
5. If you responded "no" above, can clarify which rule is insufficient? In addition, if possible, can you provide a grammatical, structural or syntactic rule that is not detailed in Appendix E.2?

E.4 Contextual accuracy Protocol

Given a table of SAE-AAE sentence pairs examples, determine whether or not their contextual accuracy is maintained.

1. As you responded "yes" a previous question,

   > *... are you capable of code-switching by alternating between SAE and AAE frequently on a daily basis in a single conversation or situation?*

   We will now provide 20 lower-cased test sentences is Table E.1.

Table E.1: SAE examples and their AAE equivalents (after using CODESWITCH).

SAE	AAE
i will go back to the house	imma go back ta da house
i don't want to go to bed	ion wanna go ta bed
he isn't my friend, but he's a king	he ain't my friend, but he a king
she is being weird to me	she been weird ta me
you all are annoying	yall annoyin
he isn't coming anymore	he ain't comin no mo
a woman is trying to walk	a woman tryna walk
this bag and that shoe are mine	dis bag n dat shoe mine
their kids are laughing	they kids laughin
john and kates have two dogs	john n kates hav two dogs
are you going through something	u goin thru sumn
what are you doing	wat r u doin
what's the temperature	wus da temperature
they have a better car than us	dey hav a betta car dan us
so you're going to the party	so your gonna go ta da party
they are singing but they can't sing	dey singing but dey can't sing
you could of have it all	u coulda hav it all
he would've had it if he was here	he woulda had it if he was here
we should have been first in line	we shoulda been first in line
he should of had the last bite	he shoulda had da last bite

2. Have you ever seen any of these words in a particular sentence in Table E.1, for example, on social media such as Twitter?

3. If you responded "yes" above, For each SAE sentence, does each plausible AAE sentence resemble adequate AAE morphological language features from a dominant AAE speaker after applying CODESWITCH?

4. If you responded "yes" above, do these pairs maintain their contextual accuracy i.e., original structure, intent, semantic equivalence and quality?

5. For dialectal (morphological and phonological) purposes, are these particular words spelt how would you say or use them? For example, texting or posting on social media?

6. If you responded "no" above, can you provide a different spelling along with its SAE equivalent?