

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Prior Definitions . . . . .	2
1.2	Research Problem . . . . .	3
1.3	Research Hypothesis . . . . .	4
1.4	Results . . . . .	4
1.5	Research Outcome . . . . .	5
1.6	Outline . . . . .	5
<b>2</b>	<b>Background and Related Work</b>	<b>7</b>
2.1	Scientific Disciplines . . . . .	8
2.1.1	Natural Language Processing . . . . .	8
2.1.2	Machine Learning . . . . .	9
2.2	Word Representations . . . . .	10
2.2.1	One Hot Representations . . . . .	10
2.2.2	Distributional Hypothesis and Distributional Representations . . . . .	11
2.2.3	Word Context Matrices . . . . .	11
2.2.4	Distributed Representations or Word Embeddings . . . . .	13
2.3	Fairness in Machine Learning . . . . .	22
2.3.1	Bias in Data . . . . .	23
2.3.2	Algorithmic Fairness . . . . .	24
2.4	Fairness in Word Embeddings . . . . .	25
2.4.1	Works on Bias Measurement in Word embeddings . . . . .	25
2.4.2	Bias Mitigation of Word Embeddings . . . . .	28
2.5	Discussion . . . . .	29
<b>3</b>	<b>WEFE Design</b>	<b>30</b>
3.1	Building Blocks . . . . .	30
3.1.1	Target Set . . . . .	31
3.1.2	Attribute Set . . . . .	31
3.1.3	Query . . . . .	31
3.1.4	Templates and Subqueries . . . . .	31
3.1.5	Fairness Metrics . . . . .	32
3.2	WEFE Ranking Process . . . . .	32
3.2.1	Creating the Scores Matrix . . . . .	33
3.2.2	Creating the Rankings . . . . .	33
3.2.3	Gathering Rankings in a Final Matrix . . . . .	33

3.3	Case Study . . . . .	34
3.3.1	Embedding models . . . . .	34
3.3.2	Queries and Query Sets . . . . .	35
3.3.3	Specific Fairness Metrics . . . . .	35
3.3.4	Results . . . . .	37
<b>4</b>	<b>WEFE Library</b>	<b>40</b>
4.1	Motivation . . . . .	41
4.2	Components . . . . .	41
4.2.1	Target and Attribute Sets . . . . .	41
4.2.2	Query . . . . .	42
4.2.3	Word Embedding Model . . . . .	42
4.2.4	Metric . . . . .	42
4.2.5	Utils . . . . .	43
4.3	Bias Measurement Processes . . . . .	43
4.3.1	Simple Query Creation and Execution . . . . .	43
4.3.2	Runners . . . . .	44
4.3.3	Aggregating Results and Calculating Rankings . . . . .	44
4.3.4	Ranking Correlations . . . . .	45
<b>5</b>	<b>Conclusions and Future Work</b>	<b>48</b>
5.1	Conclusions . . . . .	48
5.2	Future Work . . . . .	49
	<b>Bibliography</b>	<b>52</b>
	<b>Appendixes</b>	<b>57</b>
<b>A</b>	<b>Word Sets and Queries</b>	<b>57</b>
A.1	WEAT Word Sets . . . . .	57
A.2	RND Word Sets . . . . .	58
A.3	Debias Word Embeddings Word Sets . . . . .	60
A.4	Debias Multiclass Words Sets . . . . .	61
A.5	Bing Liu Sentiment Lexicon . . . . .	62
<b>B</b>	<b>Queries</b>	<b>63</b>
B.1	Gender Queries . . . . .	64
B.2	Ethnicity Queries . . . . .	65
B.3	Religion Queries . . . . .	66
<b>C</b>	<b>WEFE Library Tutorial</b>	<b>67</b>
C.1	Run a Query . . . . .	67
C.2	Running Multiple Queries . . . . .	68
C.3	Rankings . . . . .	72
C.3.1	Differences in Magnitude Using the Same Fairness Metric . . . . .	72
C.3.2	Differences in Magnitude Using Different Fairness Metrics . . . . .	73
C.3.3	Calculate Rankings . . . . .	73
C.3.4	Ranking Correlations . . . . .	76