Table of Content

1 Introduction 1

1.1 Background 1

1.2 Description of the Problem 2

1.3 Objectives 2

1.3.1 General Objectives 2

1.3.2 Specific Objectives 2

1.4 Bias Mitigation in WEFE 3

1.5 Results 3

1.6 Outline 3

2 Background and Related Work 5

2.1 Natural Language Processing 5

2.2 Word Embeddings 6

2.2.1 Tasks 6

2.2.2 Training process 6

2.2.3 Pre-trained models 7

2.2.4 Evaluation 7

2.3 Bias in Word Embeddings 8

2.3.1 Bias Measurement 8

2.3.2 Bias Mitigation 11

2.4 Word Embedding Fairness Evaluation (WEFE) 14

2.4.1 WEFE Bias Measurement Structure 15

2.4.2 WEFE Case Study 16

2.4.3 Bias Mitigation 17

3 Implemented Methods **19**

3.1 Similarities between algorithms 19

3.1.1 Gender Bias 19

3.1.2 Operations 19

3.1.3 Word Sets 20

3.1.4 Comparison between the algorithms 21

3.2 Bias Mitigation Standardization 21

3.2.1 Double Hard Debias 23

3.2.2 Half Sibling Regression 23

3.2.3 Repulsion Attraction Neutralization 24

3.2.4 Evaluation 25

3.2.5 Development Process 25

3.2.6 WEFE Repository 26

3.2.7 WEFE Documentation 26

3.3 Word Embeddings Benchmarks 26

4 Experiments **27**

4.1 Correctness check 27

4.1.1 Half Sibling Regression 27

4.1.2 Double Hard Debias 28

4.2 Performance of Algorithms and Models 29

4.2.1 Bias in word embedding models 30

4.2.2 Performance of the word embedding models 34

5 Conclusions and Future Work **38**

5.1 Conclusions 38

5.2 Future Work 39

Bibliography **42**

Annexes **43**

Annexed A Experiments Results **44**

A.1 Bias in Word Embedding Models Results 44

A.1.1 Bias Rankings 44

A.1.2 Bias Metrics 46

A.2 WEB results 58

A.3 Results including definitional pairs in DHD 59

Annexed B Links and Wordsets **64**

B.1 Links to the algorithms' original repositories 64

B.2 Definitional pairs 64

B.3 Gender Specific 64