

Conceptual and methodological problems with bias  
detection and avoidance in natural language  
processing

Alicja Dobrzeniecka

2021-06-10



# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>Cosine similarity and bias detection</b>	<b>7</b>
2.1	Word embeddings . . . . .	7
2.2	Cosine similarity and distance . . . . .	7
2.3	Cosine distance in a one-class bias detection . . . . .	7
2.4	Cosine distance in a multi-class bias detection . . . . .	7
2.5	Limitations of the approach . . . . .	7
<b>3</b>	<b>Walkthrough with the religion dataset</b>	<b>9</b>
3.1	Loading and understanding the dataset . . . . .	10
3.2	First look at the empirical distributions . . . . .	10
3.3	Looking at the islam-related words . . . . .	10
3.4	Bayesian model structure and assumptions . . . . .	10
3.5	Choosing predictors . . . . .	10
3.6	Dataset-level coefficients . . . . .	10
3.7	Model structure and assumptions . . . . .	10
3.8	Protected classes in Reddit and Google embeddings . . . . .	10
3.9	Dataset-level coefficients after debiasing . . . . .	10
3.10	Protected classes after debiasing . . . . .	10



# Chapter 1

## Introduction

Placeholder



# Chapter 2

## Cosine similarity and bias detection

Placeholder

### 2.1 Word embeddings

### 2.2 Cosine similarity and distance

### 2.3 Cosine distance in a one-class bias detection

### 2.4 Cosine distance in a multi-class bias detection

### 2.5 Limitations of the approach





## Chapter 3

# Walkthrough with the religion dataset

Placeholder

- 3.1 Loading and understanding the dataset**
- 3.2 First look at the empirical distributions**
- 3.3 Looking at the islam-related words**
- 3.4 Bayesian model structure and assumptions**
- 3.5 Choosing predictors**
- 3.6 Dataset-level coefficients**
- 3.7 Model structure and assumptions**
- 3.8 Protected classes in Reddit and Google embeddings**
- 3.9 Dataset-level coefficients after debiasing**
- 3.10 Protected classes after debiasing**