

Mitigating Gender Bias in Machine Learning Data Sets

Von

Fabian Ax, Saskia Brech, Verena Pawlas und Michelle Reiners

Inhalt

2

- ▶ Problematik
- ▶ Lösungsansatz
- ▶ Methoden
- ▶ Analyse
- ▶ ESuPol


Problematik

3

- ▶ Bias in automatisierter Sprachverarbeitung können gesellschaftliche vorhandene Bias verstärken & aufrecht erhalten
- ▶ Vor allem bei Stellenanzeigen & Rekrutierungstool großes Problem
 - ▶ Basieren auf Sprachverarbeitung / Empfehlungsalgorithmen

Lösungsansatz

4

- ▶ Framework für die Identifikation von Gender Bias in Trainingsdaten für Machine Learning
 - ▶ **Analyse** von Trainingsdaten notwendig
 - ▶ Darf keine Beeinflussung beinhalten
 - ▶ **Testen** gelernter Assoziationen
- 
- ▶ **Integration** von Fairnesskonzepten

Methoden

Korpora

5

- ▶ The British Library (BL) Korpus aus dem 19. Jahrhundert
 - ▶ Größe: > 16.000 Dokumente
 - ▶ Baseline
- ▶ The Guardian
 - ▶ Alle Artikel von 2009 - 2018

Methoden

Word Embeddings

6

- ▶ Word Embeddings
 - ▶ Word2vec
 - ▶ Trainiert auf BL Korpus
- ▶ Word lexicons
 - ▶ Basiert auf The General Inquirer Dictionary
 - ▶ Soll zeitgenössische geschlechtsspezifische Assoziationen aufzeigen, z.B.
 - ▶ Emotionen
 - ▶ Familie
 - ▶ Handeln
- ▶ Verbindung wurde mithilfe der Kosinusähnlichkeit berechnet

Methoden

Framework

7

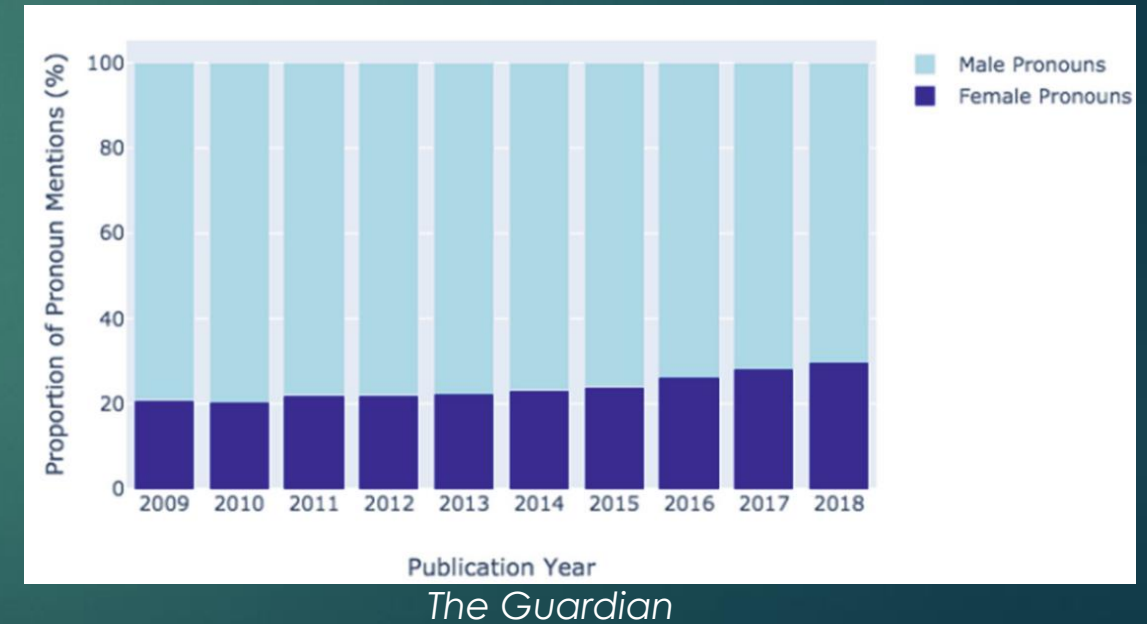
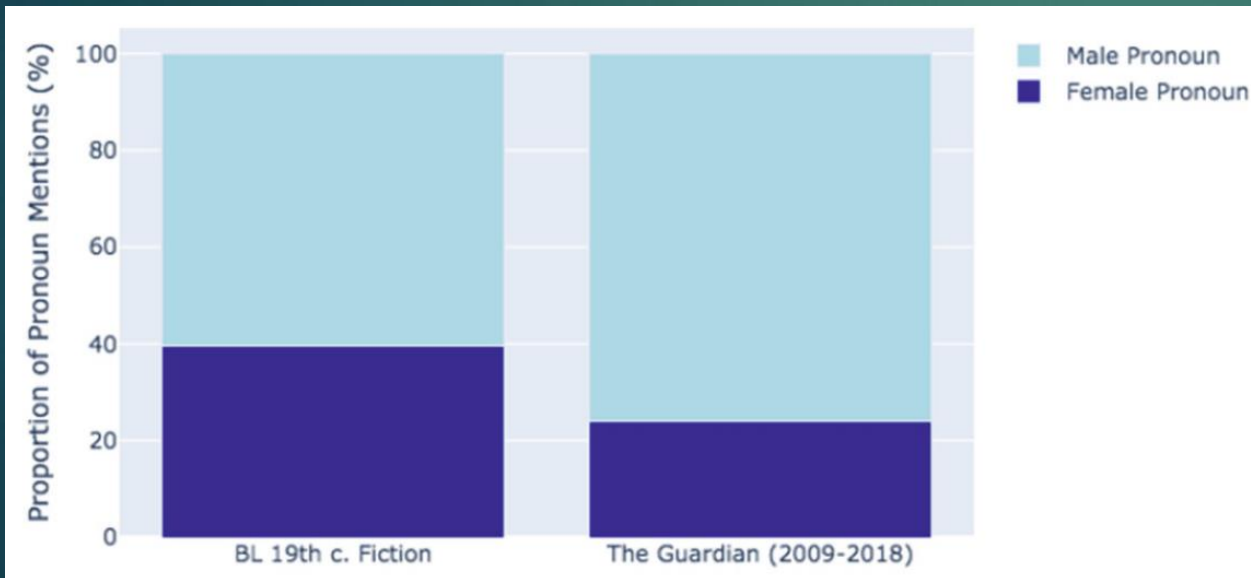
- ▶ Basiert auf
 - ▶ feministischer Kritik
 - ▶ Analyse des Sprachgebrauchs
- ▶ Hilft dabei Anzeichen für Gender Bias aufzudecken anhand
 - ▶ Präsenz von Frauen im Text
 - ▶ Geschlechterspezifische Terme
 - ▶ Modifizierte (Vor-) Terme
 - ▶ Generisches Maskulinum
 - ▶ Negative / stereotypische Assoziation

Analyse

Präsenz von Frauen im Text

8

- ▶ The Guardian hat (ausgehend von verwendeten Pronomen) höheren Gender Bias:



Analyse

Geschlechterspezifische Terme

9

- ▶ BL: **female** 2,5x häufiger verwendet als **male**
 - ▶ Male = Standardwert
 - ▶ Female = Ausnahme (muss genannt werden)
- ▶ The Guardian:
 - ▶ 2009: Benutzung auf **56%** gesunken
 - ▶ 2018: Benutzung auf **60%** gestiegen → gesteigerte Medienpräsenz

Analyse

Geschlechterspezifische Terme: Berufe

10

	Male	Female
British Library	0	28
Guardian (2018)	145	263

Premodified occupations (unique terms)

- ▶ 19. Jh.: Minimale Verwendung
 - ▶ Geschlechtsneutrale Berufe: *servant, attendant, domestic*
- ▶ Seit 2009: Anstieg der geschlechterspezifischen Terme für Berufe

Analyse

Gender und Emotionen

11

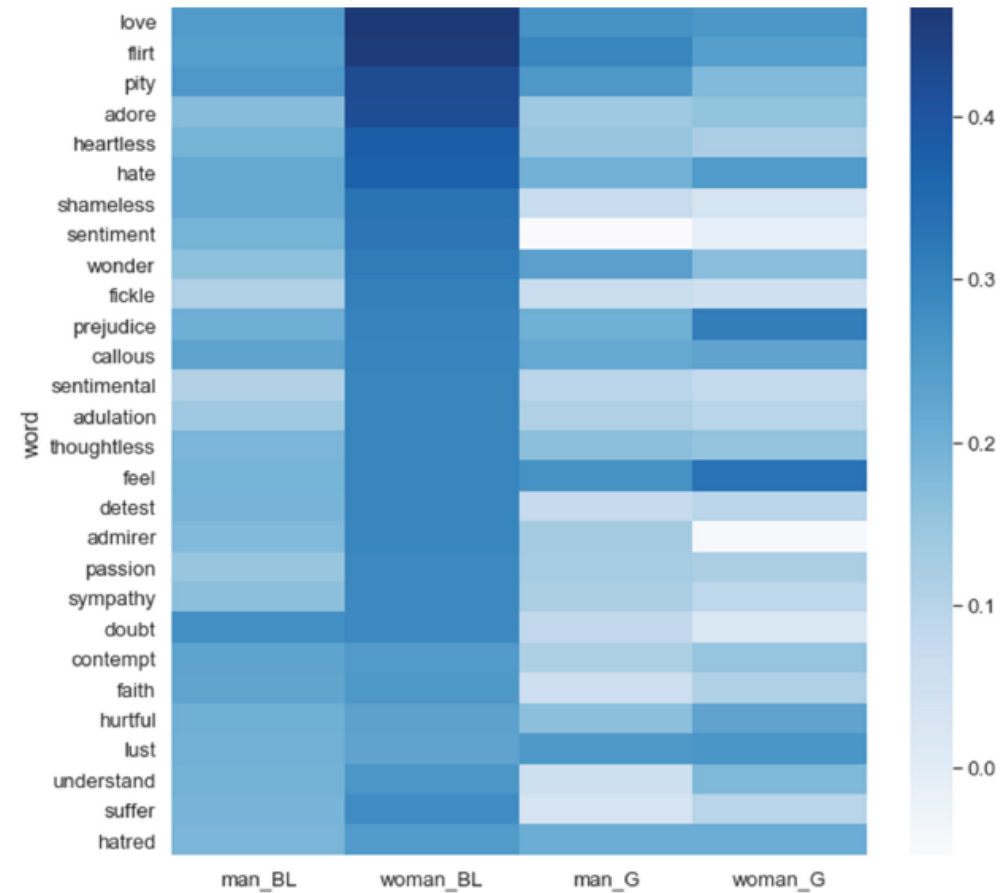


Fig. 2. Emotion: Similarity of top terms for the BL and The Guardian corpora.

Analyse

Geschlechterspezifische Handlung

12

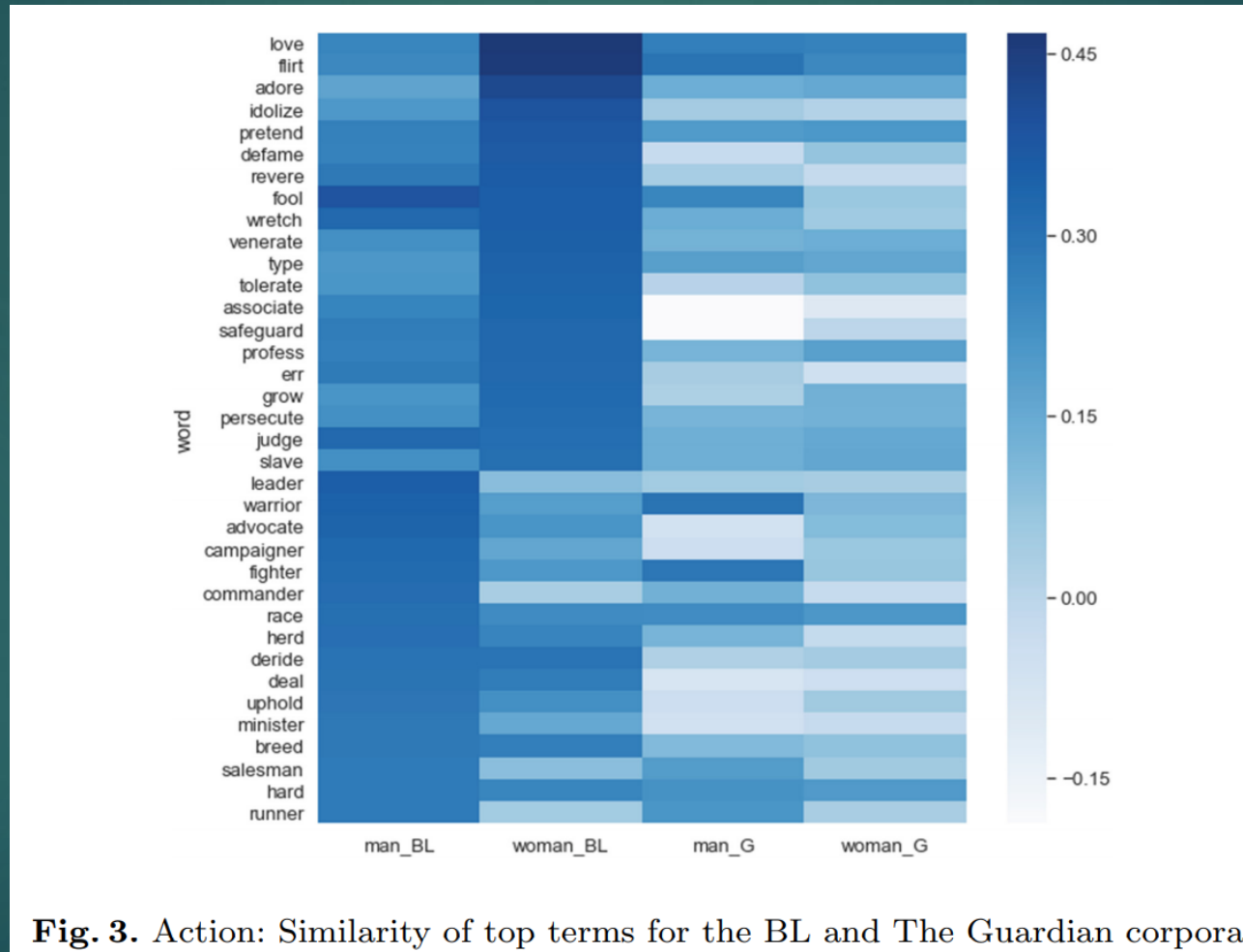


Fig. 3. Action: Similarity of top terms for the BL and The Guardian corpora.

Analyse

Geschlechterspezifische Charaktereigenschaften

13

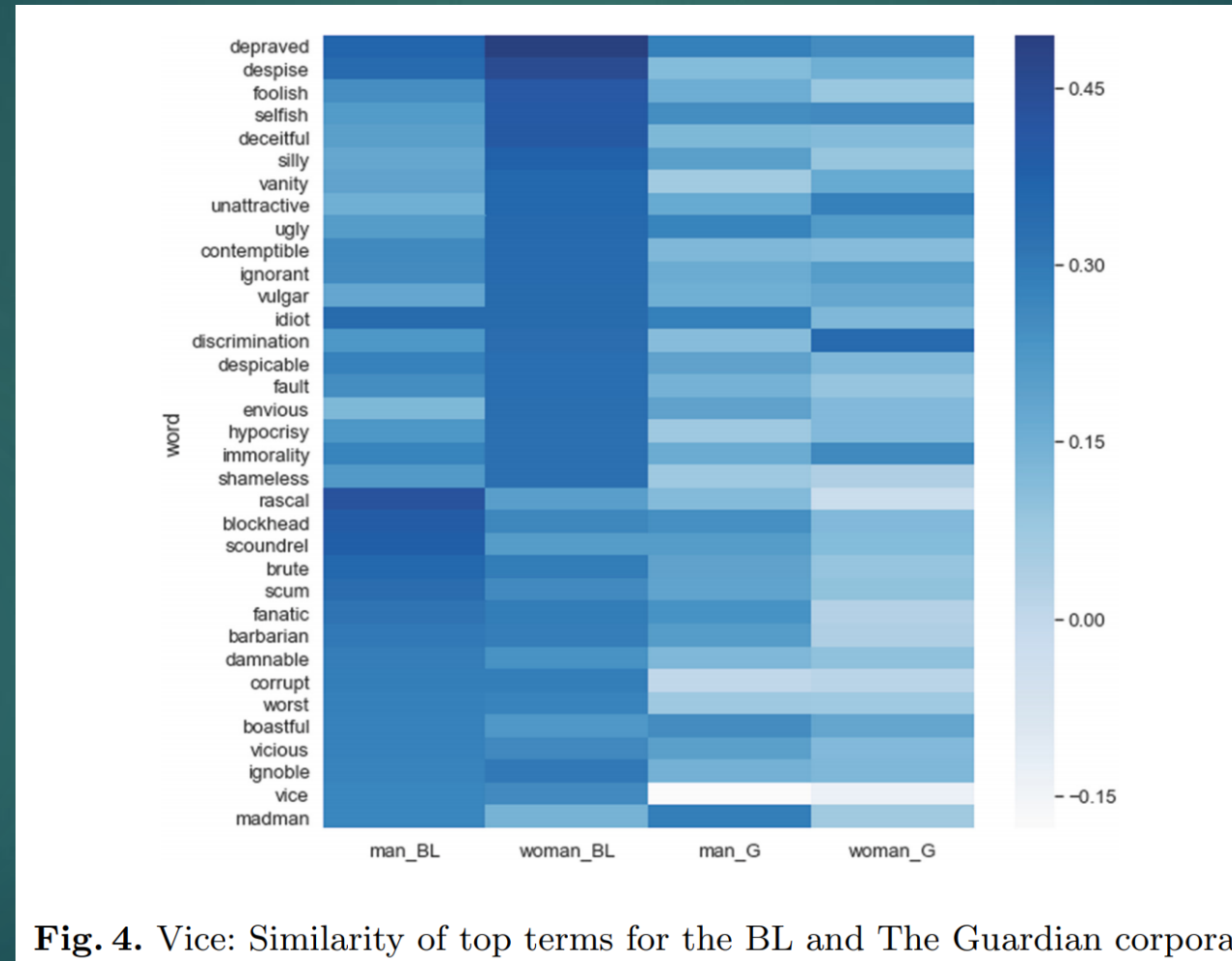


Fig. 4. Vice: Similarity of top terms for the BL and The Guardian corpora.

Analyse

Geschlechterspezifische Assoziationen mit Familie

14

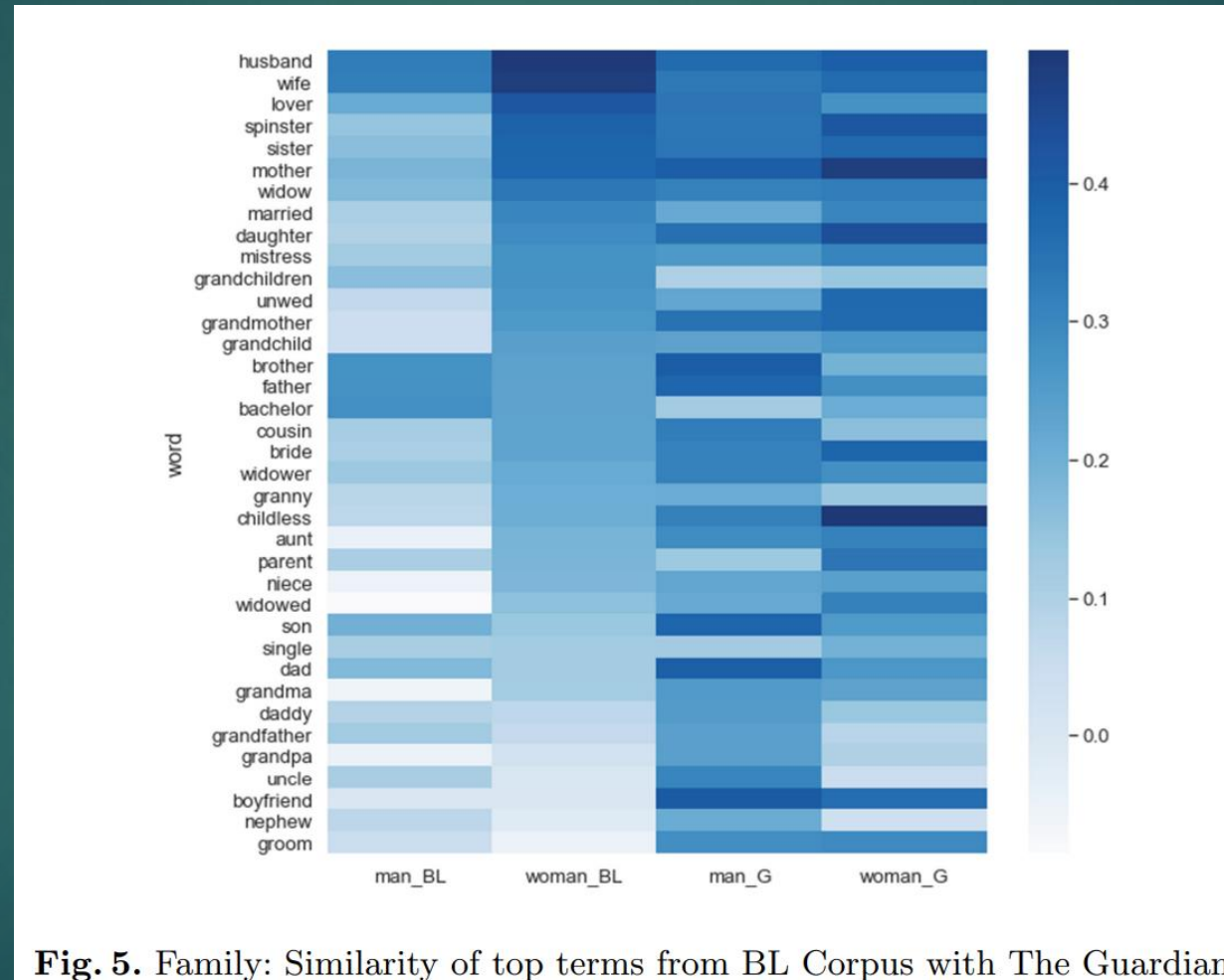


Fig. 5. Family: Similarity of top terms from BL Corpus with The Guardian.

Analyse

Reihenfolge der Binomiale

15

- ▶ „man and woman“
 - ▶ BL: in 87% der Fälle
 - ▶ The Guardian (2009): in 78% der Fälle
 - ▶ The Guardian (2018): in 74% der Fälle

- ▶ Jahre und entsprechenden Bias miteinander vergleichen
- ▶ **Baseline:** Biased Dokumentenkörper
- ▶ **Cluster-Lexikon:** Deutsches Äquivalent zu General Inquirer
 - ▶ z.B. SentiWS
- ▶ Nicht alle Ansätze des Framework können vollständig auf Datensatz/Queries angewandt werden
 - ▶ Präsenz von Frauen im Text
 - ▶ Modifizierte (Vor-) Terme

- ▶ Geschlechterspezifische Terme & Generisches Maskulinum
 - ▶ z.B. Politiker & Politikerin
 - ▶ Wird Politiker auch für Politikerin verwendet?
- ▶ Stereotypische (& negative) Assoziation
 - ▶ Bei Frauen: Schwanger, Kosmetik
 - ▶ Bei Männern: KFZ/Auto, Fußball
- ▶ **ABER** → Analyse schwierig: Gender Bias oder Personen Bias? :(

Vielen Dank für die Aufmerksamkeit!

