Automatically Neutralizing Subjective Bias in Text

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, Diyi Yang

Presentators
- Jüri Keller
- Leon Munz
- Matteo Meier

Agenda

1. Introduction
    1. Subjektiver Bias
    2. Neutralisieren von subjektivem Bias
2. Wiki Neutrality Corpus (WNC)
    1. Quelle
    2. Unterkategorien
    3. Dataset-Eigenschaften
3. Modelle
    1. Modular
    2. Concurrent
4. Performance Conclusion
5. Adaption auf ESuPol-Datensatz
    1. Neutralisieren von Bias
    2. Erkennen von Bias
6. Quellen und Links

01

Introduction

1.1 Subjektiver Bias

- Subjektiver Bias tritt auf, wenn Sprache neutral sein sollte, jedoch durch Gefühle, Meinungen oder Geschmack verzerrt wird.

Einbringen von Haltung

Wahrheitsannahme

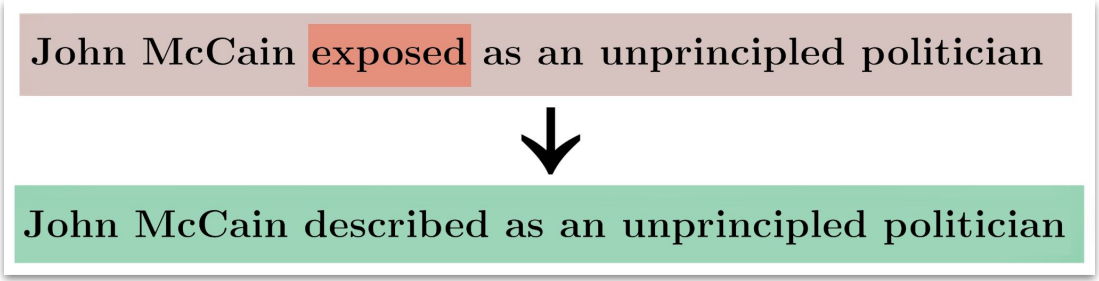
Auslösen von Zweifeln

1.2 Neutralisieren von Subjektivem Bias

- Um Subjektiven Bias zu identifizieren wird die Richtlinie für neutrale Standpunkte (NPOV) von Wikipedia genutzt.

- Diese Richtlinie beinhaltet eine Reihe von Prinzipien die unter anderen die Vermeidung der Darstellung von Meinung als Fakten aufführt.

1.2 Neutralisieren von Subjektivem Bias



Beispiel:

1. “exposed” ist ein faktisches Verb, das wie Wahrheit seines Komplements voraussetzt .

2. Das Ersetzen von “exposed” durch “described” neutralisiert den Satz.

3. Der Satz behält dabei eine ähnliche Aussage

02 —

Wiki Neutrality Corpus (WNC)

2.1 Wiki Neutrality Corpus (WNC)

- Parallelkorpus mit ca. 180.000 verzerrten und neutralisierten Satzpaaren.
- Erstellt aus Wikipedia-Edits aus der Kategorie NPOV (2004 - 2019).

2.1 Wiki Neutrality Corpus (WNC)

Drei Kategorien von subjektiver Verzerrung:

- Epistemologische Verzerrung
- Framing
- Demographische Verzerrung

Subcategory	Percent
Epistemological	25.0
Framing	57.7
Demographic	11.7

Von 500 Samples des WNC

2.1 Wiki Neutrality Corpus (WNC)

Source	Target	Subcategory
A new downtown is being developed which will bring back...	A new downtown is being developed which **which its promoters hope** will bring back..	Epistemological
The authors' **exposé** on nutrition studies	The authors' **statements** on nutrition studies	Epistemological
He started writing books **revealing** a vast world conspiracy	He started writing books **alleging** a vast world conspiracy	Epistemological
Go is **the deepest** game in the world.	Go is **one of the deepest** games in the world.	Framing
Most of the gameplay is **pilfered from** DDR.	Most of the gameplay is **based on** DDR.	Framing
Jewish forces overcome Arab **militants**.	Jewish forces overcome Arab **forces**.	Framing
A lead programmer usually spends **his career** mired in obscurity.	Lead programmers often spend **their careers** mired in obscurity.	Demographic
The lyrics are about **mankind**'s perceived idea of hell.	The lyrics are about **humanity**'s perceived idea of hell.	Demographic
Marriage is a **holy union** of individuals.	Marriage is a **personal union** of individuals.	Demographic

Table 1: Samples from our new corpus. 500 sentence pairs are annotated with "subcategory" information (Column 3).

2.1 Wiki Neutrality Corpus (WNC)

- Es besteht ein Zusammenhang zwischen dem Thema des Textes und der Häufigkeit von Subjektivem Bias.

Häufiges Auftreten von subjektivem Bias	Geringes Auftreten von subjektivem Bias
Geschichte	Meteorologie
Politik	Wissenschaft
Philosophie	Landformen
Sport	Rundfunk
Sprache	Kunstkategorien

03 —

Modelle

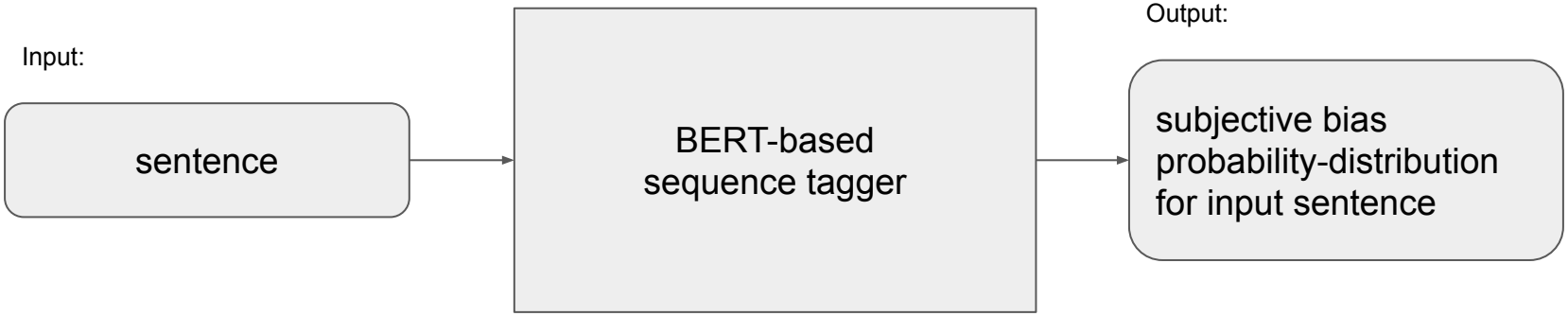
Modular

Zweistufiges Modell zur Entfernung von Subjective Bias

1. **Detection Module:**
   erkennt Wörter mit Bias.

2. **Editing Module:**
   tauscht die erkannten Wörter gegen Wörter ohne Bias.

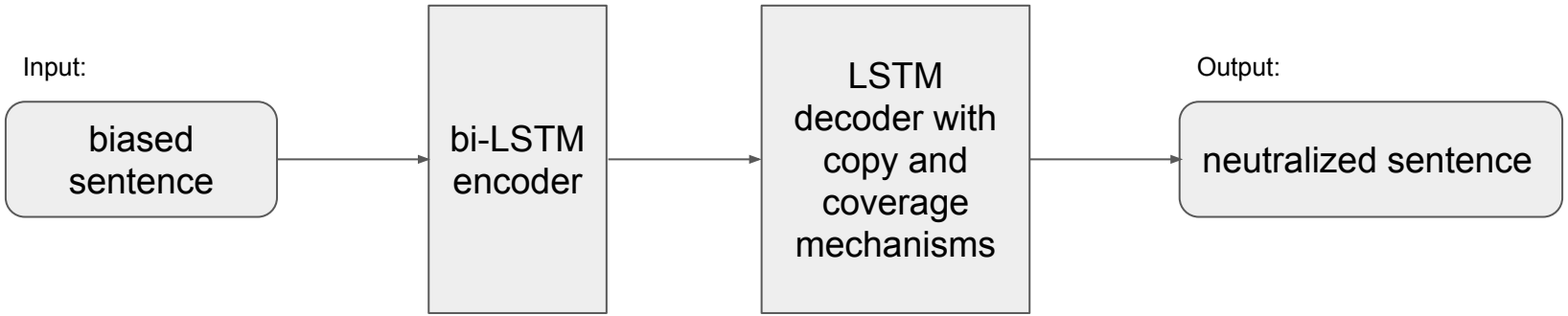
3.1 Modular: Detection Module



“Marriage is a holy union of individuals.”

“Marriage is a **holy** union of individuals.”

3.1 Modular: Editing Module



“Marriage is a **holy** union of individuals.”

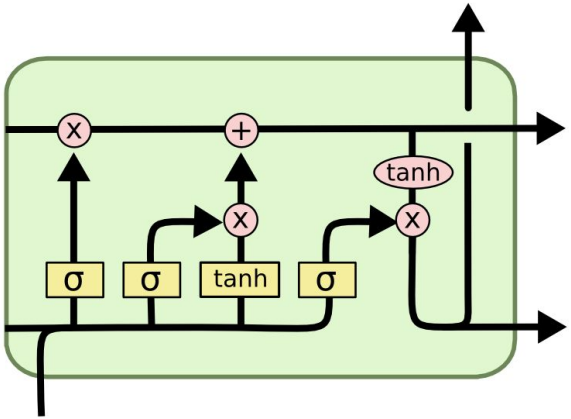
“Marriage is a **personal** union of individuals.”

3.1 Detection Module: Training

BERT-based sequence tagger:

- BERT wurde als *masked language model* trainiert. Also auf nicht gelabelten Texten.
- Um den Bias-Topic Zusammenhang zu erhalten, wird dem Input ein Topic Token hinzugefügt.
- Das Modell wird auf dem WNC Datensatz vortrainiert. Dabei haben alle Terme ein Label, ob sie Bias enthalten oder nicht.

3.1 Editing Module: Training



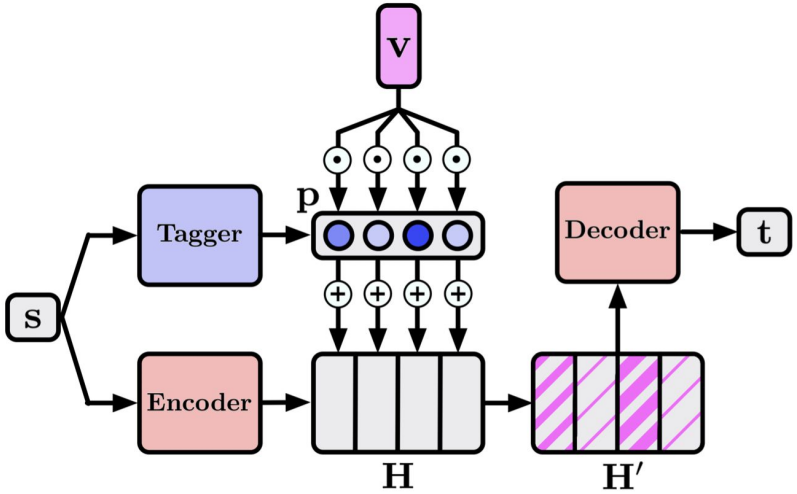
Editing Module:

- Trainiert auf dem “sauberen” Teil des WNC Datensatzes
- **bi-LSTM** bezieht vorwärts und rückwärts Beziehungen ein und encoded den Text.
- **LSTM** decodiert wieder zu Text.
- *copy mechanism* erstellt Output aus Input und neu berechneten Token
- *coverage mechanisms* verhindert wiederholungen von Wörtern und Sequenzen

3.1 Modular: Training

Final System:

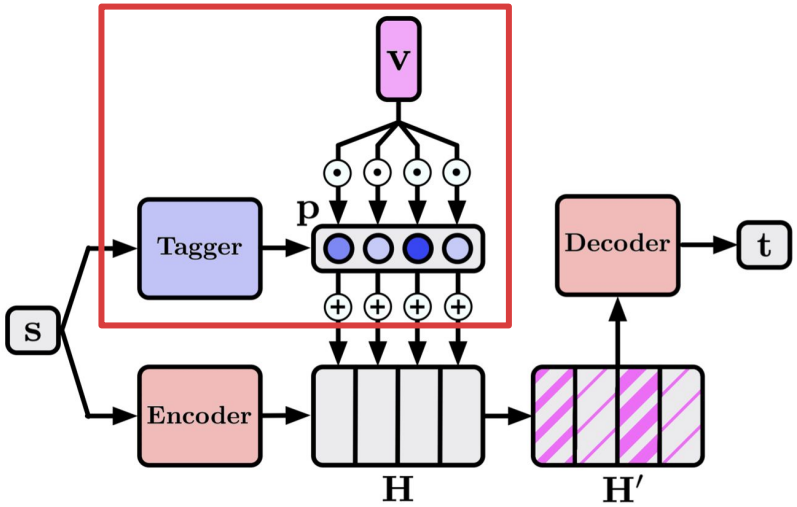
- Beide Module (Detection und Editing) sind im finalen System miteinander verbunden.
- Die *subjective bias probability-distribution* aus dem Detection Modul fließt in den Encoder mit ein.
- So wird bestimmt welche Wörter angepasst werden müssen und welche behalten werden können



3.1 Modular: Training

Final System:

- Beide Module (Detection und Editing) sind im finalen System miteinander verbunden.
- Die *subjective bias probability-distribution* aus dem Detection Modul fließt in den Encoder mit ein.
- So wird bestimmt welche Wörter angepasst werden müssen und welche behalten werden können

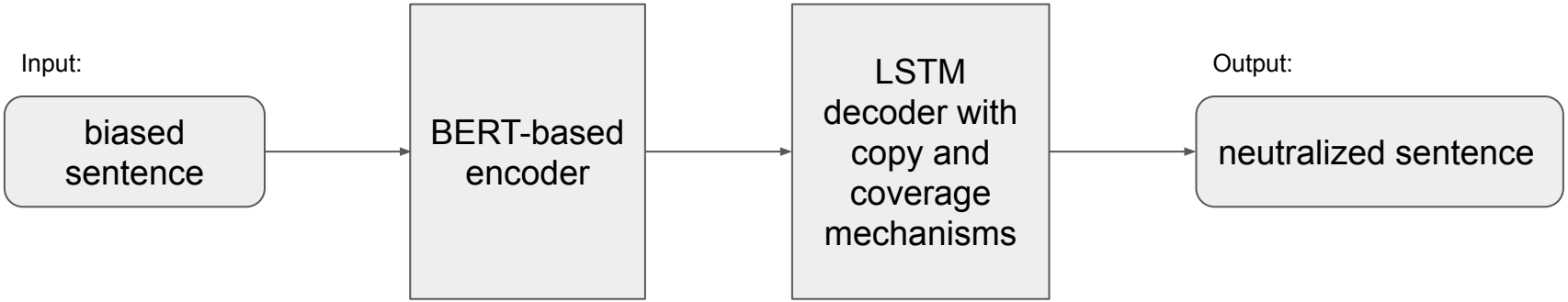


Concurrent

Einstufiges Modell zur Entfernung von Subjective Bias

Tauscht die Wörter mit Bias gegen Wörter ohne Bias.

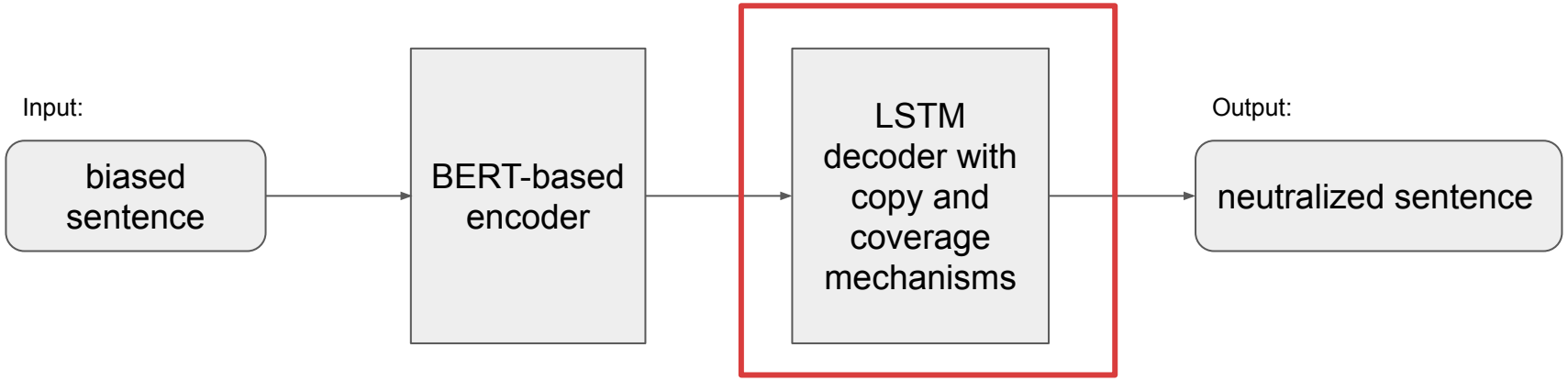
3.2 Concurrent: Editing Module



"Marriage is a **holy** union of individuals."

"Marriage is a **personal** union of individuals."

3.2 Concurrent: Editing Module



"Marriage is a **holy** union of individuals."

"Marriage is a **personal** union of individuals."

3.2 Concurrent: Training

- Ein Satz mit Subjective Bias wird direkt in einen Satz ohne Subjective Bias übersetzt.

- Anstelle des *sequence taggers* und des Encoders werden die Sätze direkt encoded und an den **LSTM** Decoder übergeben.

- Geringerer Trainingsaufwand

04

Performance Conclusion

4.1 Performance and Conclusion

Evaluierungsergebnisse bei Anwendung auf den WNC-Korpus*

Method	BLEU	Accuracy	Fluency	Bias	Meaning
Source Copy	91.33	0.00	-	-	-
Detector (always delete biased word)	92.43*	38.19*	-0.253*	-0.324*	1.108*
Detector (predict substitution from biased word)	92.51	36.57*	-0.233*	-0.327*	1.139*
Delete Retrieve (ST) (Li et al. 2018)	88.46*	14.50*	-0.209*	-0.456*	1.294*
Back Translation (ST) (Prabhumoye et al. 2018)	84.95*	9.92*	-0.359*	-0.390*	1.126*
Transformer (MT) (Vaswani et al. 2017)	86.40*	24.34*	-0.259*	-0.458*	0.905*
Seq2Seq (MT) (Luong, Pham, and Manning 2015)	89.03*	23.93	-0.423*	-0.436*	1.294*
Base	89.13	24.01	-	-	-
+ *loss*	90.32*	24.10	-	-	-
+ *loss* + *pretrain*	92.89*	34.76*	-	-	-
+ *loss* + *pretrain* + *detector* (MODULAR)	93.52*	**45.80***	-0.078	**-0.467***	0.996*
+ *loss* + *pretrain* + *BERT* (CONCURRENT)	**93.94**	44.87	**0.132**	-0.423*	**0.758***
Target copy	100.0	100.0	-0.077	-0.551*	1.128*

- **hoher manueller Aufwand** bei der Erstellung der Quality Measures
- **Modular** ist stärker bei der **Reduzierung des Bias**
- **Concurrent** hat bessere Werte für die **Erhaltung Fluency und Meaning**

Bestätigung der WNC-Evaluierung-Ergebnisse durch Evaluierung auf anderen Korpussen (IBC Corpus, News Headlines, Trump Speeches)

05 —

Adaption auf den ESuPol-Datensatz

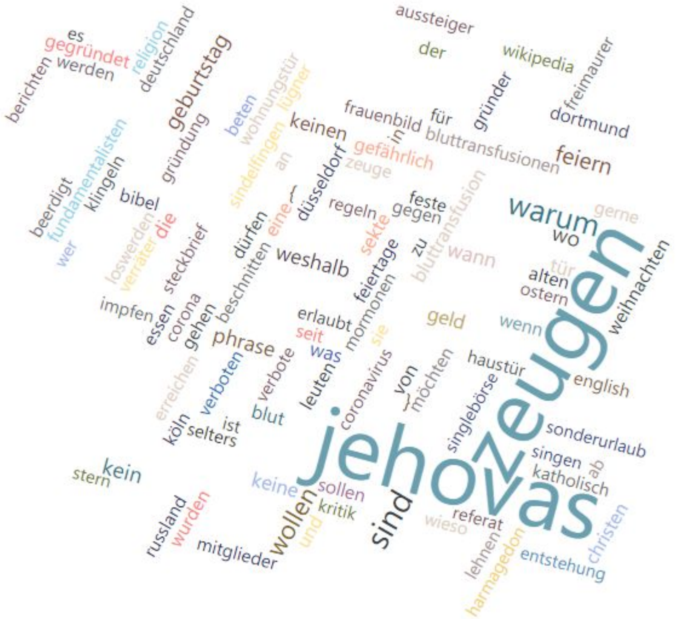
5.1 Adaption des Neutralisierens von Bias

mögliche Adaptionen:

- Modular/Concurrent Editing Module zur Neutralisierung von Suggestion Terms in Suchmaschinen (Aufsetzen einer “Bias-freien” Suchmaschine)
- Subjective Bias bereinigte Vergleiche/Statistiken
  - Hypothesentest auf signifikante Sprach-Unterschiede zwischen der Suche im Kontext ethnischer vs religiöser Minorities

5.1.1 Neutralisieren von Bias - Beispiel 1

This plot contains unprocessed data from ESuPol “suggestions_minorities_slice.csv”.



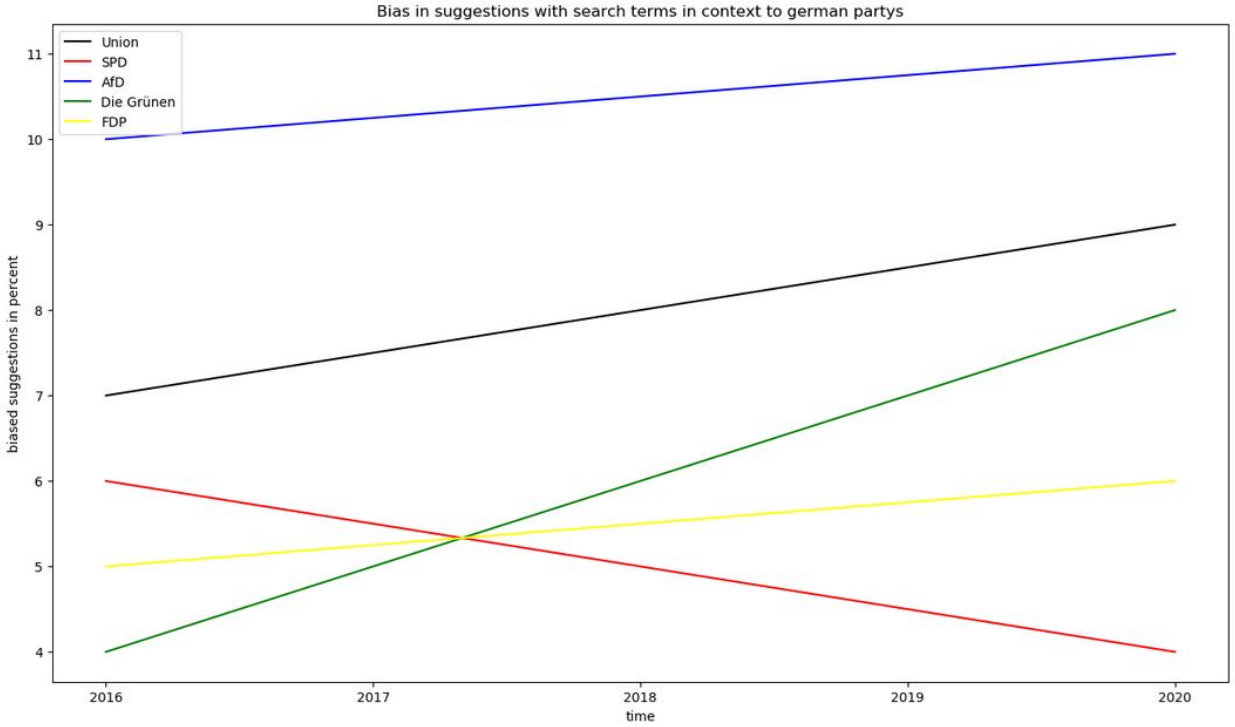
Differenzbetrachtung **neutralisierter Suchvorschläge** hinsichtlich ethnischer/religiöser Minderheiten

5.2 Adaption des Erkennens von Bias

mögliche Adapationen:

- Modular Detection Module zur Erkennung von Bias
  - Bewertung des Gesellschaftsbildes für Politiker aus bestimmten Parteien
  - Bias-Scores für Suchmaschinen/Länder
  - Erweiterung mit WNC-Bias-Subcategory (Einschränkung: Kategorisierung wurde nur auf ein 500er-Sample des WNC vorgenommen)

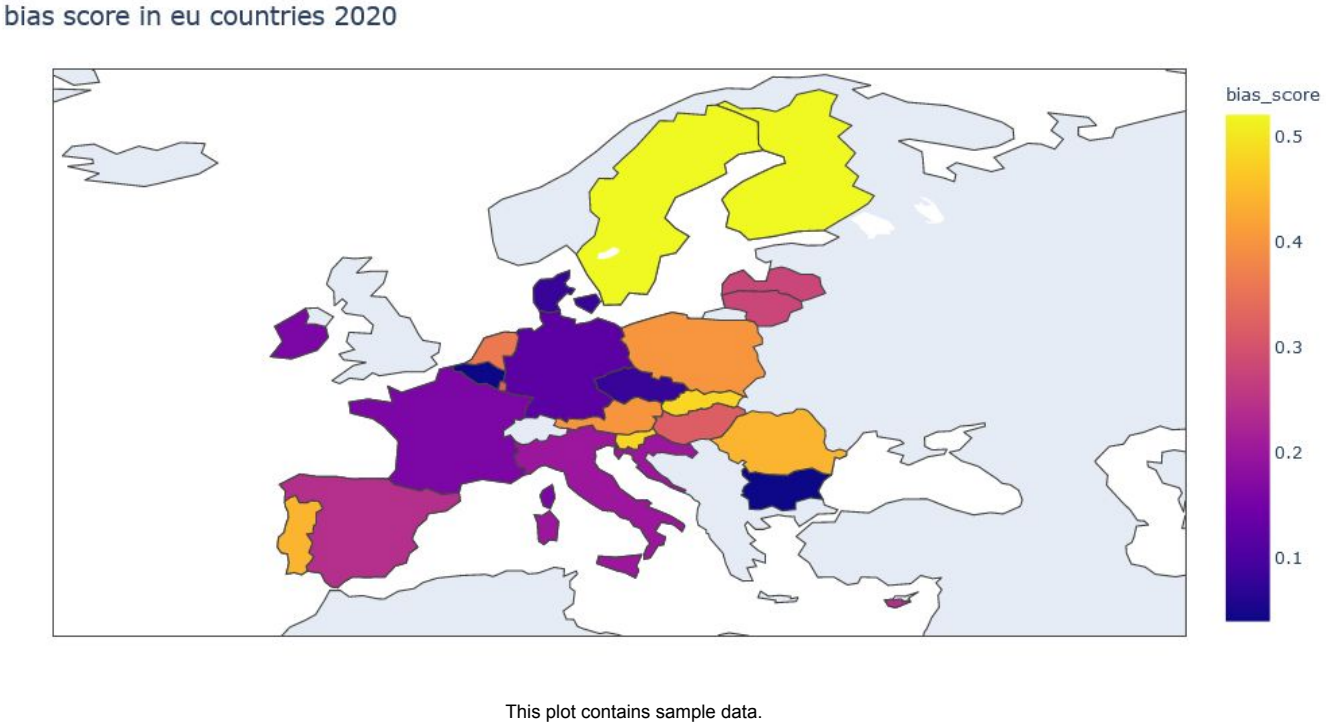
5.2.1 Erkennen von Bias - Beispiel 1



This plot contains sample data.

Bias Bestimmung bei Suchanfragen in Zusammenhang mit Politikern bestimmter Parteien

5.2.2 Erkennen von Bias - Beispiel 2



Bias Score für verschiedene Länder im ESuPol Datensatz

06 —

Quellen und Links

6 Quellen und Links

Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., & Yang, D. (2020).
Automatically neutralizing subjective bias in text.
In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, No. 01, pp.480-489).
https://arxiv.org/abs/1911.09709

Original Code und Daten auf GitHub: https://github.com/rpryzant/neutralizing-bias

Suggestions Minorities Slice from ESuPol Dataset: https://github.com/irgroup/dis25-2021

Sämtliche Diagramme und Darstellungen sind entweder aus dem o.g. Paper übernommen oder stammen aus eigener Darstellung.