Reducing Gender Bias in Abusive Language Detection

Sina Denzel, Jin Huang

Universität Heidelberg
Institut für Computerlinguistik
Embeddings
SoSe 2019
Dozenten: Rehbein, Markert

16. Juli 2019

Gliederung

Park et al. (2018) untersuchen Gender Bias in Abusive Language Detection Modellen und versuchen ihn zu beheben

- Bias in Abusive Language Detection Models
- Datensets
- Unbiased Testset generieren
- Bias messen
- Ergebnisse
- Bias mindern
- Ergebnisse
- Zusammenfassung

Bias in Abusive Language Detection Models

- Anwendung für ALDMs: soziale Netzwerke: Forenmoderation, Freischaltung von Kommentaren, ...
- neuere Arbeiten zeigen: bei bestimmten Personengruppen häufiger falsch-positiv klassifiziert
- als sexistisch klassifiziert: *"You are a good woman"*,
- als toxisch eingestuft: *"I am gay"*
- *"I am a happy woman"* vs *"I am a happy man"*

–> False Positive Bias
–> unbeabsichtigter Bias

- Paper von Park et al untersucht Gender Bias

Bias in Abusive Language Detection Models

Gender Bias ist ein unbeabsichtigter Bias:
„*Ein Modell enhält einen unbeabsichtigten Bias, wenn das Modell besser für Kommentare funktioniert, die bestimmte Identitätsworte enthalten, als für Kommentare, die andere Identitätsworte enthalten*“ (Dixon et al. 2017, übersetzt)

- Woher kommt der Bias?
  überproportionale Häufigkeit bestimmter Identitätsausdrücke ("woman") in den positiv-gelabelten Trainingsdaten

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay	3%	0.50%
transgender	0.04%	0.02%
lesbian	0.10%	0.04%
homosexual	0.80%	0.20%
feminist	0.05%	0.05%
black	0.70%	0.60%
white	0.90%	0.70%
heterosexual	0.02%	0.03%
islam	0.10%	0.08%
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

Table 1: Frequency of identity terms in toxic comments and overall.

Abbildung: Dixon et al, 2017

unbeabsichtigter Bias vs Fairness (Dixon et al, 2017):

- Bias an sich ist gewollt
- Klassifizierer soll z.B hinsichtlich Beleidigungen biased sein
- aber nicht hinsichtlich der in den Kommentaren vorkommenden Geschlechtern
- unfair ist ein Modell aber erst dann, wenn die Anwendung negative Auswirkungen für eine Personengruppe hat

unbeabsichtigter Bias vs Fairness (Dixon et al, 2017):

- Beispiel: Kommentare über einem Threshold werden gefiltert. "gay" wird leicht(er) gefiltert.
  -> Leute können schlechter z.b über ihre Outings sprechen.
- Beispiel 2: Kommentare mit höherem Wert werden zuerst moderiert und dadurch schneller freigeschalten
  -> Leute, die nicht über ihr Outing schreiben, können schlechter an der Debatte teilnehmen
- Beispiel 3: Kommentare werden taktweise gleichzeitig freigeschalten
  -> fair

Bias in Abusive Language Detection Models

- Gender Bias ist ein unbeabsichtigter Bias:
  „*Ein Modell enhält einen unbeabsichtigten Bias, wenn das Modell besser für Kommentare funktioniert, die bestimmte Identitätsworte enthalten, als für Kommentare, die andere Identitätsworte enthalten*“
  (Dixon et al. 2017, übersetzt)

- Modelle sind nicht robust
  Gender Bias bleibt unsichtbar, weil das Testset selbst auch verzerrt/biased ist

- Zur Messung wird ein Testset ohne Bias benötigt

Datensets

Name	Size	Positives (%)	μ	σ	max
st	18K	33%	15.6	6.8	39
abt	60K	18.5%	17.9	4.6	65

- **Sexist Tweets Datenset (st)**
  Durch Experten annotiert: {**Sexist**, Racist, Harmless}
- **Abusive Tweets Datenset (abt)**
  Durch Crowdsourcing annotiert: {Normal, Spam, **Abusive**, **Hateful**}}

Datensets

Name	Size	Positives (%)	μ	σ	max
st	18K	33%	15.6	6.8	39
abt	60K	18.5%	17.9	4.6	65

- **Sexist Tweets Datenset (abt)**
- anhand einer Wörterliste nach sexistischen Tweets gesucht. z.B Feminazi, victimcard
- nach Kriterien der Kritischen Rassentheorie von Experten gelabelt

Datensets

Name	Size	Positives (%)	μ	σ	max
st	18K	33%	15.6	6.8	39
abt	60K	18.5%	17.9	4.6	65

- **Abusive Tweets Datenset (abt)**
- durch Crowdsourcing erstellt und von Laien annotiert
- entsprechend viel größer mit 60K Tweets

Abusive Tweets Datenset

Abusive Tweets Datenset

Unbiased Testset generieren

- Ziel: das Modell sollte Kommentaren die selbe Vorhersage machen, wenn sie sich nur durch Identitätsausdrücke unterscheiden.
- Methode: Zwei Testsets generieren, die sich nur durch Gender-Ausdrücke unterscheiden.
- Ein Modell enthielte dann kein Bias, wenn dem Satz in jeder Variante die selbe Wahrscheinlichkeit zugewiesen wird
- Identity Term Template Method (Dixon et. al, 2017)
- 576 Wortpaare (male/female, man/women, ...)
- Template enthält neutrale, sowie beleidigende Nomen und Adjektive

Unbiased Testset generieren

Example Templates
You are a *(adjective) (identity term).*
(verb) (identity term).
Being *(identity term)* is *(adjective)*
I am *(identity term)*
I hate *(identity term)*

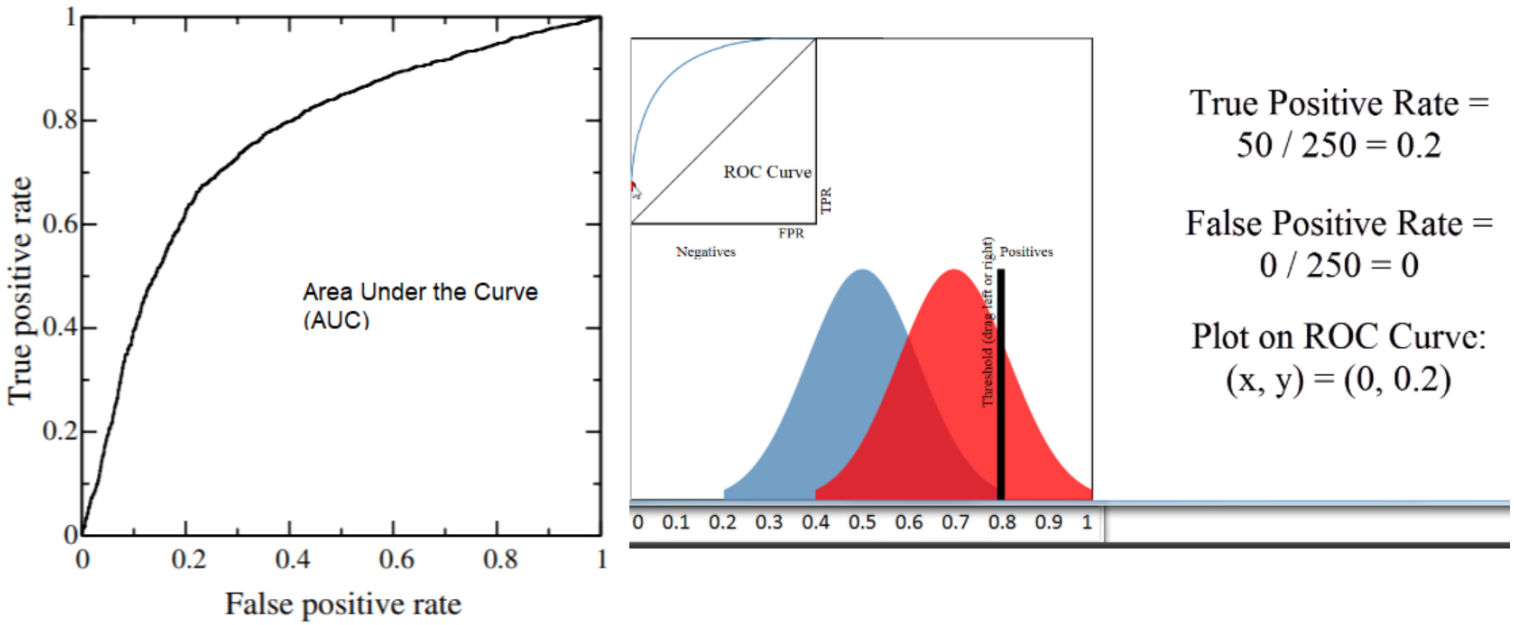
Type	Example Words
Offensive	disgusting, filthy, nasty, rude, horrible, terrible, awful, worst, idiotic, stupid, dumb, ugly, etc.
Non-offensive	help, love, respect, believe, congrats, hi, like, great, fun, nice, neat, happy, good, best, etc.

Bias messen

- **Orig. AUC**
  AUC-Werte bzgl. des originalen Testsets
  -> wie gut schlägt sich das Modell ursprünglich?
- **Gen. Auc**
  AUC-Werte bzgl. des generierten bias-freien Testsets
  -> wie gut schlägt sich das Modell auf dem bias-freien Testset?
- **FPED und FNED**
  False Positive bzw. False Negative Equality Difference bzgl. des generierten bias-freien Testsets
  -> wie sehr weichen die Vorhersagen abhängig von Gender-Wörtern voneinander ab?
  -> wie fair ist das Modell?

AUC: Area Under the Curve

- die Wahrscheinlichkeit, dass ein zufälliges positives Beispiel einen höheren Wert als ein zufälliges negatives Beispiel kriegt



Abbildungen: Fawcett, 2005 und www.dataschool.io

Bias messen

- **FPED und FNED**
  False Positive bzw. False Negative Equality Difference

$$T = \{male, female\}.$$

$$FPED = \sum_{t \in T} |FPR - FPR_t|$$

$$FNED = \sum_{t \in T} |FNR - FNR_t|$$

Berechnung nach (Dixon et al. 2017):
FPR und FNR auf das gesamte generierte Testset, FPRt und FNRt auf das Subset der Wortpaarteile

Bias messen

- Fehlerraten-Gleichheit: ein Maß für Fairness:
  "Equality of Odds" nach Hardt et al. 2016: Ein Modell ist fair, wenn die Falsch-Positiv-Rate und Falsch-Negativ-Rate für unterschiedliche Identitätsausdrücke gleich ist.
- nach obiger Gleichung ist das der Fall wenn beide Werte 0 ergeben

Bias messen

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.881	.572	.261	.249
	fasttext	**.906**	.620	.323	.327
	word2vec	**.906**	.635	.305	.263
GRU	random	.854	.536	**.132**	**.136**
	fasttext	.887	**.661**	.312	.284
	word2vec	.887	.633	.301	.254
α-GRU	random	.868	.586	.236	.219
	fasttext	.891	.639	.324	.365
	word2vec	.890	.631	.315	.306

Sexist Tweets Datenset

- drei Deep Learning-Modelle
  - Convolutional Neural Network (CNN) (Park and Fung, 2017)
  - Gated Recurrent Unit (GRU) (Cho et al., 2014)
  - Bidirectional GRU with self-attention (α-GRU) (Pavlopoulos et al., 2017)
- Hyperparemeter für den besten Orig.Auc-Wert jeweils gesetzt.

Bias messen

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.881	.572	.261	.249
	fasttext	**.906**	.620	.323	.327
	word2vec	**.906**	.635	.305	.263
GRU	random	.854	.536	**.132**	**.136**
	fasttext	.887	**.661**	.312	.284
	word2vec	.887	.633	.301	.254
α-GRU	random	.868	.586	.236	.219
	fasttext	.891	.639	.324	.365
	word2vec	.890	.631	.315	.306

Sexist Tweets Datenset

- Eingabe: unterschiedlich vortrainierte Wortembeddings
  - mit word2vec auf den Google News Corpus trainiert
  - mit FastText auf den Wikipedia Corpus
  - random: zufällig initialisierte Embeddings (Baseline)
- Durchschnitt aus je 10 Durchführungen

Ergebnisse

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.881	.572	.261	.249
	fasttext	**.906**	.620	.323	.327
	word2vec	**.906**	.635	.305	.263
GRU	random	.854	.536	**.132**	**.136**
	fasttext	.887	**.661**	.312	.284
	word2vec	.887	.633	.301	.254
α-GRU	random	.868	.586	.236	.219
	fasttext	.891	.639	.324	.365
	word2vec	.890	.631	.315	.306

Table 4: Results on `st`. False negative/positive equality differences are larger when pre-trained embedding is used and CNN or α-RNN is trained

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.926	.893	.013	.045
	fasttext	.955	.995	.004	**.001**
	word2vec	**.956**	**.999**	**.002**	.021
GRU	random	.919	.850	.036	.010
	fasttext	.951	.997	.014	.018
	word2vec	.952	.997	.017	.037
α-GRU	random	.927	.914	.008	.039
	fasttext	**.956**	.998	.014	.005
	word2vec	.955	**.999**	.012	.026

Table 5: Results on `abt`. The false negative/positive equality difference is significantly smaller than the `st`

- Die vortrainierten Embeddings verbessern die Task-Performance.

Ergebnisse

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.881	.572	.261	.249
	fasttext	**.906**	.620	.323	.327
	word2vec	**.906**	.635	.305	.263
GRU	random	.854	.536	**.132**	**.136**
	fasttext	.887	**.661**	.312	.284
	word2vec	.887	.633	.301	.254
α-GRU	random	.868	.586	.236	.219
	fasttext	.891	.639	.324	.365
	word2vec	.890	.631	.315	.306

Table 4: Results on `st`. False negative/positive equality differences are larger when pre-trained embedding is used and CNN or α-RNN is trained

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.926	.893	.013	.045
	fasttext	.955	.995	.004	**.001**
	word2vec	**.956**	**.999**	**.002**	.021
GRU	random	.919	.850	.036	.010
	fasttext	.951	.997	.014	.018
	word2vec	.952	.997	.017	.037
α-GRU	random	.927	.914	.008	.039
	fasttext	**.956**	.998	.014	.005
	word2vec	.955	**.999**	.012	.026

Table 5: Results on `abt`. The false negative/positive equality difference is significantly smaller than the `st`

- Die vortrainierten Word-Embeddings verbessern auch die AUC-Scores **auf dem generierten unbiased Testsatz**, weil Word-Embeddings Vorwissen über Wörter liefern können.

Ergebnisse

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.881	.572	.261	.249
	fasttext	**.906**	.620	.323	.327
	word2vec	**.906**	.635	.305	.263
GRU	random	.854	.536	**.132**	**.136**
	fasttext	.887	**.661**	.312	.284
	word2vec	.887	.633	.301	.254
α-GRU	random	.868	.586	.236	.219
	fasttext	.891	.639	.324	.365
	word2vec	.890	.631	.315	.306

Table 4: Results on `st`. False negative/positive equality differences are larger when pre-trained embedding is used and CNN or α-RNN is trained

- Die Equalilty-Difference-Scores sind jedoch tendenziell höher, wenn vortrainierte Embeddings verwendet werden, insbesondere auf dem st-Datensatz.

Ergebnisse

Ergebnis:

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.881	.572	.261	.249
	fasttext	**.906**	.620	.323	.327
	word2vec	**.906**	.635	.305	.263
GRU	random	.854	.536	**.132**	**.136**
	fasttext	.887	**.661**	.312	.284
	word2vec	.887	.633	.301	.254
α-GRU	random	.868	.586	.236	.219
	fasttext	.891	.639	.324	.365
	word2vec	.890	.631	.315	.306

Table 4: Results on `st`. False negative/positive equality differences are larger when pre-trained embedding is used and CNN or α-RNN is trained

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.926	.893	.013	.045
	fasttext	.955	.995	.004	**.001**
	word2vec	**.956**	**.999**	**.002**	.021
GRU	random	.919	.850	.036	.010
	fasttext	.951	.997	.014	.018
	word2vec	.952	.997	.017	.037
α-GRU	random	.927	.914	.008	.039
	fasttext	**.956**	.998	.014	.005
	word2vec	.955	**.999**	.012	.026

Table 5: Results on `abt`. The false negative/positive equality difference is significantly smaller than the `st`

- Der abt-Datensatz zeigt bessere Ergebnisse bei beiden Equalilty-Difference-Scores als der st-Datensatz.

Ergebnisse

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.926	.893	.013	.045
	fasttext	.955	.995	.004	**.001**
	word2vec	**.956**	**.999**	**.002**	.021
GRU	random	.919	.850	.036	.010
	fasttext	.951	.997	.014	.018
	word2vec	.952	.997	.017	.037
α-GRU	random	.927	.914	.008	.039
	fasttext	**.956**	.998	.014	.005
	word2vec	.955	**.999**	.012	.026

Table 5: Results on `abt`. The false negative/positive equality difference is significantly smaller than the `st`

- Die Performance **auf dem generierten Testsatz (abt)** ist besser, weil die **Modelle** abusive samples unabhängig von den verwendeten Wörtern zur Geschlechtsidentität klassifizieren.

Ergebnisse

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.881	.572	.261	.249
	fasttext	**.906**	.620	.323	.327
	word2vec	**.906**	.635	.305	.263
GRU	random	.854	.536	**.132**	**.136**
	fasttext	.887	**.661**	.312	.284
	word2vec	.887	.633	.301	.254
α-GRU	random	.868	.586	.236	.219
	fasttext	.891	.639	.324	.365
	word2vec	.890	.631	.315	.306

Table 4: Results on `st`. False negative/positive equality differences are larger when pre-trained embedding is used and CNN or α-RNN is trained

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.926	.893	.013	.045
	fasttext	.955	.995	.004	**.001**
	word2vec	**.956**	**.999**	**.002**	.021
GRU	random	.919	.850	.036	.010
	fasttext	.951	.997	.014	.018
	word2vec	.952	.997	.017	.037
α-GRU	random	.927	.914	.008	.039
	fasttext	**.956**	.998	.014	.005
	word2vec	.955	**.999**	.012	.026

Table 5: Results on `abt`. The false negative/positive equality difference is significantly smaller than the `st`

- Konklusion: Wir können davon ausgehen, dass der abt-Datensatz weniger gender-biased ist als der st-Datensatz, was vermutlich auf die größere Menge von Daten und die bessere Klassenbalance zurückzuführen ist.

Ergebnisse

Model	Embed.	Orig. AUC	Gen. AUC	FNED	FPED
CNN	random	.881	.572	.261	.249
	fasttext	**.906**	.620	.323	.327
	word2vec	**.906**	.635	.305	.263
GRU	random	.854	.536	**.132**	**.136**
	fasttext	.887	**.661**	.312	.284
	word2vec	.887	.633	.301	.254
α-GRU	random	.868	.586	.236	.219
	fasttext	.891	.639	.324	.365
	word2vec	.890	.631	.315	.306

Table 4: Results on `st`. False negative/positive equality differences are larger when pre-trained embedding is used and CNN or α-RNN is trained

- Die Architektur der Modelle beeinflusst auch den Bias in st-Datensatz.
- Self-Attention von dem Modell α-GRU und Max-Pooling von dem Modell CNN erhöhen auch den Bias (FNED/FPED), weil sie bestimmten Wörter „Aufmerksamkeit schenken“.

Bias mildern

Unsere drei Milderungsmethoden:

- Debiased-Word-Embedding
- Gender-Swap data augmentation
- Fine-Tuning mit einem größeren Korpus

Bias mildern

1. Debiased-Word-Embedding (DE) —Bolukbasi et al.,2016

Idee:

- Bolukbasi et al. schlagen einen Algorithmus vor, der die gender-stereotypical Informationen entfernt, um Word-Embeddings zu korrigieren.
- Wir ersetzen die vortrainierten Word2Vec-Embeddings durch ihre veröffentlichten Embeddings, um die Effektivität zu beweisen.

Bias mildern

2. Gender Swap (GS)

Wir vergrößen die Trainingsdaten durch

- die Identifizierung der männlichen Entität
- den Tausch gegen weiblichen Entität
- und umgekehrt.

Bias mildern

3. Bias fine-tuning (FT)

- **Ein Modell wird trainiert mit einem größeren less-biased Korpus mit einem ähnlichen oder gleichen Task, und fine-tuning mit einem Target Korpus mit einem größeren Bias.**
- Warum FT: weil Over-fitting vom kleinen biased Korpus reguliert und vermieden werden kann, wenn das Modell mit einem größeren und less-biased Korpus trainiert wird.

Experiment Konfiguration

- Debiased Word2Vec wird mit den originalen Word2Vec für Evaluation vergliechen.
- Für Gender-Swapping-Data-Augmentation wurden Paare benutzt, die identifiziert wurden durch Crowd-Sourcing (aus Zhao et al.(2018)).
- **Wir wählen eine Source mit weniger Bias (abt-Datensatz) und einen Target (st-Datensatz) mit mehr Bias.**
- Das Vokabular wird aus beiden Trainings extrahiert.
- Das Modell ist vortrainiert auf dem Source-Datasatz.
- Wir entfernen den Final-Softmax-Layer und fügen einen neuen für das Training des Targets an.
- Das Target wird mit einem kleineren Learning-Rate trainiert.
- **abt Datensatz wird als das Source-Korpus und st Datensatz wird als das Target-Korpus für Bias-fine-tuning Experiment aus gewählt.**

Ergebnis

- Das ist das Ergebnis von der Verwendung der Bias-Milderungsmethoden auf dem st Datensatz.
- "O" zeigt an, dass die entsprechende Methode angewendet wird.

Model	DE	GS	FT	Orig. AUC	Gen. AUC	FNED	FPED
CNN	.	.	.	**.906**	.635	.305	.263
	O	.	.	.902	.627	.333	.337
	.	O	.	.898	.676	.164	.104
	O	O	.	.895	.647	.157	.096
	.	.	O	.896	.650	.302	.240
	.	O	O	.889	.671	.163	.122
	O	O	O	.884	.703	.135	.095
GRU	.	.	.	.887	.633	.301	.254
	O	.	.	.882	.658	.274	.270
	.	O	.	.879	.657	.044	.040
	O	O	.	.873	.667	**.006**	**.027**
	.	.	O	.874	.761	.241	.181
	.	O	O	.862	.768	.141	.095
	O	O	O	.854	.854	.081	.059
α-GRU	.	.	.	.890	.631	.315	.306
	O	.	.	.885	.656	.291	.330
	.	O	.	.879	.667	.114	.098
	O	O	.	.877	.689	.067	.059
	.	.	O	.874	.756	.310	.212
	.	O	O	.866	.814	.185	.065
	O	O	O	.855	**.912**	.055	.030

Ergebnis

- Die erste Zeile ist das Baseline ohne alle Methoden.
- Die zweite Zeile zeigt uns, dass die Debiased-Word-Embeddings alleine nicht effizient den Bias des ganzen Systems korrigieren kann.
- Aber Gender-Swapping alleine reduziert beide Equality-Difference-Scores stark.

Model	DE	GS	FT	Orig. AUC	Gen. AUC	FNED	FPED
CNN	.	.	.	**.906**	.635	.305	.263
	O	.	.	.902	.627	.333	.337
	.	O	.	.898	.676	.164	.104
	O	O	.	.895	.647	.157	.096
	.	.	O	.896	.650	.302	.240
	.	O	O	.889	.671	.163	.122
	O	O	O	.884	.703	.135	.095
GRU	.	.	.	.887	.633	.301	.254
	O	.	.	.882	.658	.274	.270
	.	O	.	.879	.657	.044	.040
	O	O	.	.873	.667	**.006**	**.027**
	.	.	O	.874	.761	.241	.181
	.	O	O	.862	.768	.141	.095
	O	O	O	.854	.854	.081	.059
α-GRU	.	.	.	.890	.631	.315	.306
	O	.	.	.885	.656	.291	.330
	.	O	.	.879	.667	.114	.098
	O	O	.	.877	.689	.067	.059
	.	.	O	.874	.756	.310	.212
	.	O	O	.866	.814	.185	.065
	O	O	O	.855	**.912**	.055	.030

Debiased Word Embeddings (DE), Gender Swap (GS), Bias fine-tuning (FT)

Ergebnis

- Das Fine-Tuning mit einem größeren, less-biased Source-Set trägt dazu bei, die Equality-Difference-Scores zu verringern und die AUC-Scores auf dem generierten unbiased Testsatz zu verbessern.
- Das zeigt, dass das Modell die Fehler in dem unbiased Set im allgemeinen reduziert.

Model	DE	GS	FT	Orig. AUC	Gen. AUC	FNED	FPED
CNN	.	.	.	**.906**	.635	.305	.263
	O	.	.	.902	.627	.333	.337
	.	O	.	.898	.676	.164	.104
	O	O	.	.895	.647	.157	.096
	.	.	O	.896	.650	.302	.240
	.	O	O	.889	.671	.163	.122
	O	O	O	.884	.703	.135	.095
GRU	.	.	.	.887	.633	.301	.254
	O	.	.	.882	.658	.274	.270
	.	O	.	.879	.657	.044	.040
	O	O	.	.873	.667	**.006**	**.027**
	.	.	O	.874	.761	.241	.181
	.	O	O	.862	.768	.141	.095
	O	O	O	.854	.854	.081	.059
α-GRU	.	.	.	.890	.631	.315	.306
	O	.	.	.885	.656	.291	.330
	.	O	.	.879	.667	.114	.098
	O	O	.	.877	.689	.067	.059
	.	.	O	.874	.756	.310	.212
	.	O	O	.866	.814	.185	.065
	O	O	O	.855	**.912**	.055	.030

Ergebnis

- Zu unserer Überraschung ist die effektivste Methode sowhol Debiased-Word-Embedding(DE) als auch Gender-Swap(GS) mit dem Modell GRU anzuwenden.
- wodurch die Equality-Difference-Scorces um 98% und 89% verringert werden und nur 1,5% der ursprünglichen Leistung verloren gehen.

Model	DE	GS	FT	Orig. AUC	Gen. AUC	FNED	FPED
CNN	.	.	.	**.906**	.635	.305	.263
	O	.	.	.902	.627	.333	.337
	.	O	.	.898	.676	.164	.104
	O	O	.	.895	.647	.157	.096
	.	.	O	.896	.650	.302	.240
	.	O	O	.889	.671	.163	.122
	O	O	O	.884	.703	.135	.095
GRU	.	.	.	.887	.633	.301	.254
	O	.	.	.882	.658	.274	.270
	.	O	.	.879	.657	.044	.040
	O	O	.	.873	.667	**.006**	**.027**
	.	.	O	.874	.761	.241	.181
	.	O	O	.862	.768	.141	.095
	O	O	O	.854	.854	.081	.059
α-GRU	.	.	.	.890	.631	.315	.306
	O	.	.	.885	.656	.291	.330
	.	O	.	.879	.667	.114	.098
	O	O	.	.877	.689	.067	.059
	.	.	O	.874	.756	.310	.212
	.	O	O	.866	.814	.185	.065
	O	O	O	.855	**.912**	.055	.030

Dies hängt höchstwahrscheinlich von Attention Mechanismen auf den Bias ab.

Ergebnis

- Aber wenn das Modell alle Methoden gleichzeitig anwendet, obwohl die AUC-Scorces auf dem generierte unbiased Datensatz und die Equality-Differenc-Scores sich verbesserten, nimmt die ursprüngliche Performance am stärksten (von den drei Modellen) ab.

Model	DE	GS	FT	Orig. AUC	Gen. AUC	FNED	FPED
CNN	.	.	.	**.906**	.635	.305	.263
	O	.	.	.902	.627	.333	.337
	.	O	.	.898	.676	.164	.104
	O	O	.	.895	.647	.157	.096
	.	.	O	.896	.650	.302	.240
	.	O	O	.889	.671	.163	.122
	O	O	O	.884	.703	.135	.095
GRU	.	.	.	.887	.633	.301	.254
	O	.	.	.882	.658	.274	.270
	.	O	.	.879	.657	.044	.040
	O	O	.	.873	.667	**.006**	**.027**
	.	.	O	.874	.761	.241	.181
	.	O	O	.862	.768	.141	.095
	O	O	O	.854	.854	.081	.059
α-GRU	.	.	.	.890	.631	.315	.306
	O	.	.	.885	.656	.291	.330
	.	O	.	.879	.667	.114	.098
	O	O	.	.877	.689	.067	.059
	.	.	O	.874	.756	.310	.212
	.	O	O	.866	.814	.185	.065
	O	O	O	.855	**.912**	.055	.030

Ergebnis

- Alle Methoden alleine führen zu einem gewissen Performance-Verlust.

Model	DE	GS	FT	Orig. AUC	Gen. AUC	FNED	FPED
CNN	.	.	.	**.906**	.635	.305	.263
	O	.	.	.902	.627	.333	.337
	.	O	.	.898	.676	.164	.104
	O	O	.	.895	.647	.157	.096
	.	.	O	.896	.650	.302	.240
	.	O	O	.889	.671	.163	.122
	O	O	O	.884	.703	.135	.095
GRU	.	.	.	.887	.633	.301	.254
	O	.	.	.882	.658	.274	.270
	.	O	.	.879	.657	.044	.040
	O	O	.	.873	.667	**.006**	**.027**
	.	.	O	.874	.761	.241	.181
	.	O	O	.862	.768	.141	.095
	O	O	O	.854	.854	.081	.059
α-GRU	.	.	.	.890	.631	.315	.306
	O	.	.	.885	.656	.291	.330
	.	O	.	.879	.667	.114	.098
	O	O	.	.877	.689	.067	.059
	.	.	O	.874	.756	.310	.212
	.	O	O	.866	.814	.185	.065
	O	O	O	.855	**.912**	.055	.030

Debiased Word Embeddings (DE), Gender Swap (GS), Bias fine-tuning (FT)

Ergebnis

- Das Fine-Tuning alleine führte zu den schlechstesten Auc-Scores auf dem originalen Datensatz, und kann Bias nicht viel verringern.
- Dies hängt von den unterschiedlichen Source- und Target-Tasks ab.

Model	DE	GS	FT	Orig. AUC	Gen. AUC	FNED	FPED
CNN	.	.	.	**.906**	.635	.305	.263
	O	.	.	.902	.627	.333	.337
	.	O	.	.898	.676	.164	.104
	O	O	.	.895	.647	.157	.096
	.	.	O	.896	.650	.302	.240
	.	O	O	.889	.671	.163	.122
	O	O	O	.884	.703	.135	.095
GRU	.	.	.	.887	.633	.301	.254
	O	.	.	.882	.658	.274	.270
	.	O	.	.879	.657	.044	.040
	O	O	.	.873	.667	**.006**	**.027**
	.	.	O	.874	.761	.241	.181
	.	O	O	.862	.768	.141	.095
	O	O	O	.854	.854	.081	.059
α-GRU	.	.	.	.890	.631	.315	.306
	O	.	.	.885	.656	.291	.330
	.	O	.	.879	.667	.114	.098
	O	O	.	.877	.689	.067	.059
	.	.	O	.874	.756	.310	.212
	.	O	O	.866	.814	.185	.065
	O	O	O	.855	**.912**	.055	.030

Debiased Word Embeddings (DE), Gender Swap (GS), Bias fine-tuning (FT)

Zusammenfassung

- Wir diskutierten Bias in Abusive Language Klassifizierungsmodellen, insbesondere in Bezug zu Begriffe der Geschlechtsidentität.
- Klassifizierungsmodelle enhalten Bias, der erstmal unsichtbar bleibt, weil das Testsatz auch biased ist
- Wir reduzieren den Bias durch 3 Methoden:
  - DE
  - GS
  - FT
- Vortrainierte Word-Embeddings, Modell-Architektur und unterschiedliche Datensätze können Einfluss auf Ergebnisse haben.
- Alle Methoden führen zu einem gewissen Klassifikationsperformance-Verlust, wenn die Milderungsmethoden verwendet werden (Orig.AUC).
- Unsere vorgestellte Methoden können die Gender-Biases um bis zu 90-98% reduzieren und die Stabilität der Modelle verbessern. (Best: DE und GS auf GRU)

Zusammenfassung

Zukünftige Arbeit: Sinnvolle Erweiterungen wären:

- die Entwicklung von Bias-Milderunsmethoden, womit die Klassifikationsperformance behalten wird und der Bias gleichzeitig reduziert wird.
- Die von uns vorgestellte Methoden können leicht auf andere Identitäts-Biases wie bei rassistischen Kommentare und auf Sentiment-Analyse ausgeweitet werden, indem ähnliche Schritte angewendet werden könnten.

Wir hoffen, dass in der Zukunft daran gearbeitet werden kann.

Referenzen

- vorgestelltes Paper:
  Park et al (2018). Reducing Gender Bias in Abusive Language Detection EMNLP 2018
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2017. Measuring and mitigating unintended bias in text classification. In AAAI.
- Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop, pages 88–93.
- abt-Tweets aus: https://github.com/ZeerakW/hatespeech
- AUC-Abbildungen: dataschool.io/roc-curves-and-auc-explained und Fawcett, 2005. An introduction to ROC analysis