Wenn Algorithmen diskriminierend handeln

Johanna Zitt

Universität Leipzig

26.11.2020

Übersicht

- Einige Beispiele
  - Gender Bias
  - Gesichtserkennung
- Gründe für diskriminierende Algorithmen
- Lösungsansätze
- Gesellschaftlicher Diskurs
- Diskussion

Beispiel 1: Risikobewertung

- Risk score zur Einschätzung der Rückfallquoten von Straftäter*Innen
- Schwarzen Personen wird ein höheres Risiko als bei weißen Personen bei ähnlichen Straftaten berechnet.
- Vorhersage Wahrscheinlichkeit bei 61%

[1] Angwin, J.; Larson J.; Mattu S.; Kirchner L.: *Machine Bias: There´s software used across the country to predict duture criminals. And it´s biased against black.*, in ProPublica, 23.04.2016

Beispiel 1: Risikobewertung

- Algorithmus hat kein Wissen über die Hautfarbe der Personen.
- Risiko wird anhand von Fragenkatalog berechnet.
- Private Unternehmen entwickelt Bewertungs Software.

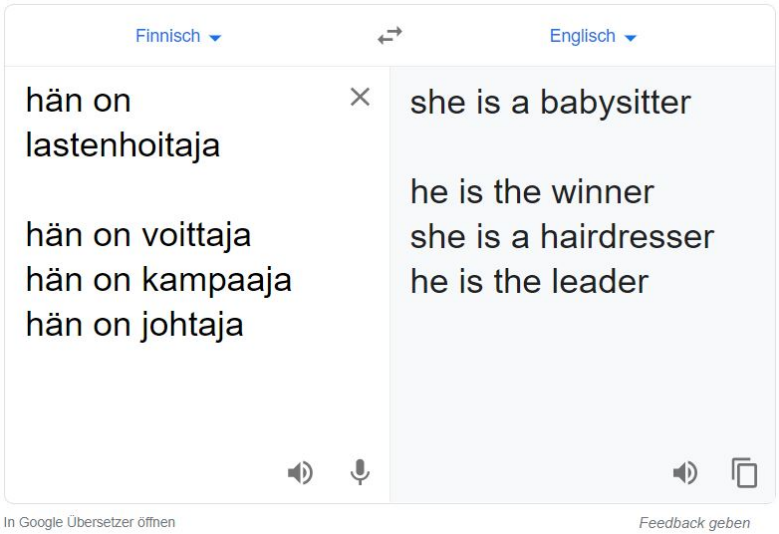
41. How many of your friends/acquaintances are gang members?
☐ None ☑ Few ☐ Half ☐ Most

42. How many of your friends/acquaintances are taking illegal drugs regularly (more than a couple times a month)?
☑ None ☐ Few ☐ Half ☐ Most

[1] Angwin, J.; Larson J.; Mattu S.; Kirchner L.: *Machine Bias: There´s software used across the country to predict duture criminals. And it´s biased against black.*, in ProPublica, 23.04.2016

Beispiel 2: Gender Bias

Googel Translate übersetzt Geschlechter neutrale Sätze in Geschlechter spezifische Sätze einer anderen Sprache.

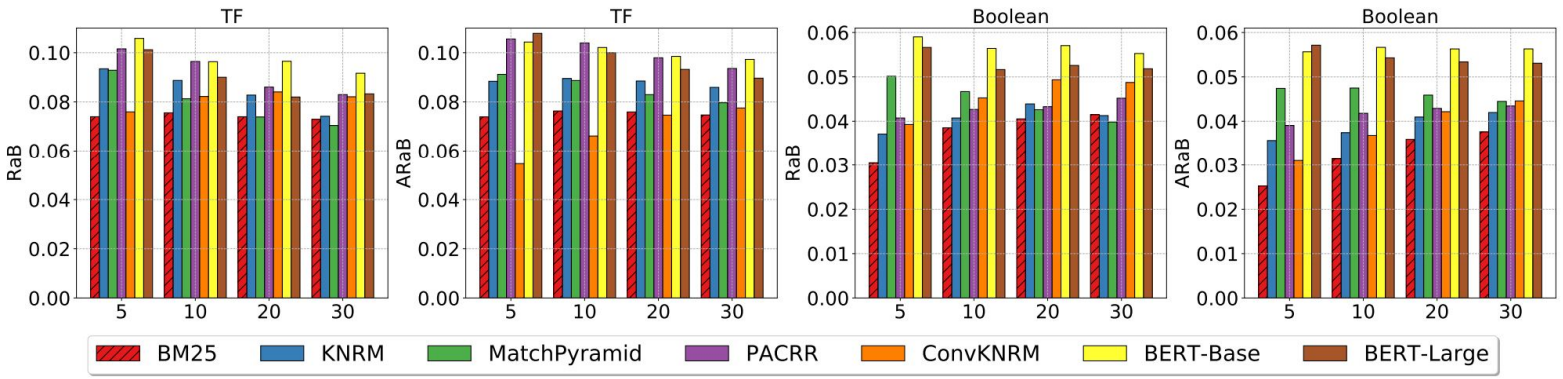


Beispiel 2: Gender Bias in IR

- Untersuchungen von Gender Bias in Information Retrieval
- Testen von 7 verschiedenen IR Modellen: BM25 und verschiedene neuronale Ranking Modelle.
- Metriken geben an wie häufig "he", "him", "man" oder "male" im Text erhalten waren im Vergleich zu weiblichen Gegenstücken.
- Nutzten von geschlechtsneutralen queries.

[2] Rekabsaz, N.; Schedl M.: *Do Neural Ranking Models Intensify Gender Bias?*, In Proceedings of 43rd ACM SIGIR Conference, 15.06.2020

Beispiel 2: Gender Bias in IR



- Alle IR Modelle haben Bias zu männlich geprägten Texten.
- Alle neuronalen Modelle zeigten einen stärkeren männlichen Bias als BM25.

[2] Rekabsaz, N.; Schedl M.: *Do Neural Ranking Models Intensify Gender Bias?*, In Proceedings of 43rd ACM SIGIR Conference, 15.06.2020

Beispiel 2: Benachteiligung aufgrund des Geschlechts

Frauen werden bei Apple-Card bei der Beurteilung Ihrer Zahlungsfähigkeit benachteiligt und bekommen bis zu 10 mal weniger Kreditrahmen als Männer bei gleichem Vermögen.

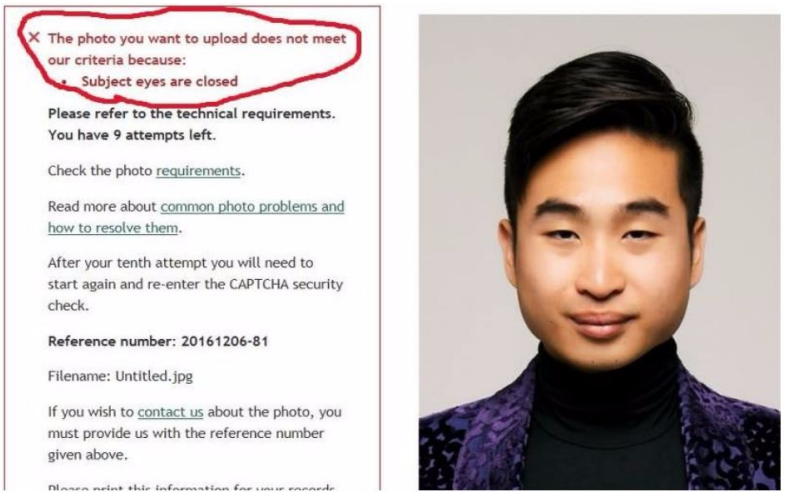
Amazons Vorauswahl von künftiger Mitarbeiter*Innen wurde nach vier jähriger Testphase eingestellt, weil Programm klar männliche Bewerber bevorzugte.

[3] Evers, A.: *Sexistische Algorithmen: Apple beanchteiligt weibliche Kunden.* in eRecht24, 17.11.2019.

[4] Redaktion: *Sexistische Algorithmen: Amazons künstliche Intelligenz zur Bewerberauswahl benachteiligte systematisch Frauen.* in Meedia, 17.10.2018.

Beispiel 3: Gesichtserkennung

Neuseeländische Reisepassbeantragung akzeptiert Bild nicht, da Augen angeblich geschlossen sind.



[5] Rentzsch, F.: *Computer-Panne oder Diskriminierung? Warum das Passfoto dieses Asiaten abgelehnt wurde*, in Buisness Insider, 9.12.2016

Beispiel 3: Gesichtserkennung

- Geschlechterklassifizierung von IBM, Micosoft und Face++ wurden auf ihre Performance hinsichtlich Geschlecht und Hautfarbe getestet.
- Datensatz von Parlamentsmitgliedern (44.89% weiblich und 47% Person of Colour)
- Bilder von schwarzen Frauen schnitten deutlich am schlechtesten ab.

	PoC/M	PoC/F	W/M	W/F
Microsoft	94.0	**79.2**	100	98.3
Face++	99.3	**65.5**	99.2	94.0
IBM	88.0	**65.3**	99.7	92.9

[6] Buolamwini, J. A.: *Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers*, 10.09.2017

Algorithmen und Software können diskriminieren.

Diskriminierung durch Algorithmen sind keine einzelnen Fälle, sondern treten häufig auf.

Aber woran liegt das?

Grund 1: Bias in Trainingsdaten

„Stecken wir Vorurteile rein, kommen Vorurteile raus“[7]

-Margaret Mitchel-

Trainingsdaten mit denen man künstliche Neuronale Netze trainiert, beinhalten bereits einen Bias, den die Software lernt und dann reproduziert.

Neuronale Netze reproduzieren so auch Vorurteile und Diskriminierung.

[7] Wolfgangel, E.: *Programmierter Rassismus: Woher Algorithmen ihre Vorurteile haben – und warum die so gefährlich sind.* in Zeit Online. (19.06.2018)

Grund 2: Data Gap

Algorithmen lernen das, was sie als Trainingsdaten erhalten.

Fehlende Diversität in Trainingsdaten kann zu fehlerhaftem Verhalten führen.

Grund 3: Wenig Diversität und wenig diverse Tests

Wenig Diversität in Unternehmen führt zu wenig Aufmerksamkeit auf benachteiligte Personengruppen.

Software wird selten auf bestimmte Bias geprüft.

[8] Leavy, S.: *Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning*, 28.05.2018.

Probleme

„Like all technologies before it, AI will reflect the values of its creators“[9]

- Kate Crowford -

- Personen, die Software entwickeln, bilden nicht die Gesellschaft ab.
- private Unternehmen legen Algorithmen nicht offen
  $\rightarrow$ für Forschung und Gesellschaft nicht zugänglich
- Neuronale Netze arbeiten als eine Blackbox und sind schwer zu durchschauen.

[9] Crawford, K.: *Artificial Intelligence´s White Guy Problem*, in New York Times, 25.06.2018.

Lösungsansatz: Neuronale Netze verstehen

- Deep Neural Networks werden in Geschlechter Klassifikationsaufgabe näher betrachtet.
- Ziel: verstehen welche Bildpunkte der Algorithmus zur Erkennung nutzt.
- Layer-wise Relevance Propagation, um Relevanz der einzelnen Pixel zu berechnen.
- Die Algorithmen konzentrieren sich bei Frauen mehr auf Haaransatz und Augen.
- Bei Männern mehr auf untere Gesichtshälfte.

[10] Lapuschkin, S.; Binder A.; Müller, K.; Samek W.: *Understanding and Comparing Deep Neural Networks for Age and Gender Classification*, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

Lösungsansatz: Neuronale Netze verstehen

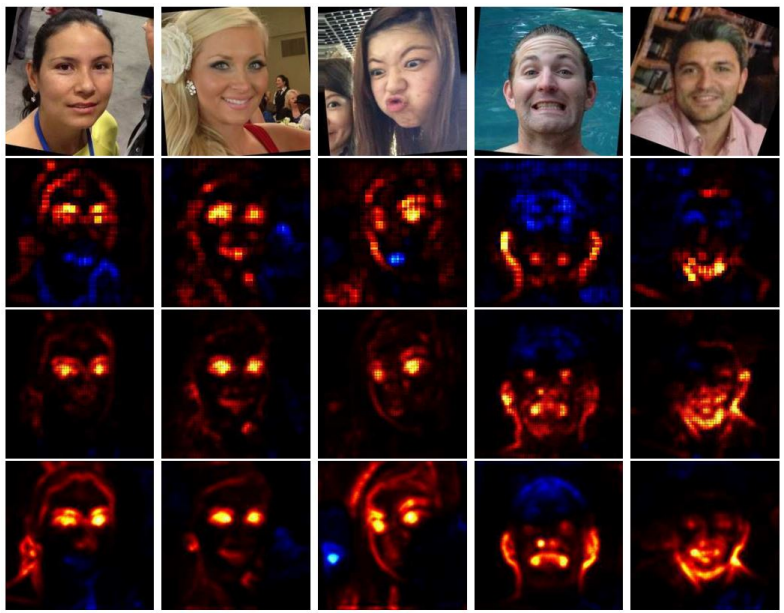


Figure 1: Inputbilder, Wärmekarte für CaffeNet, GoogleNet und VGG-16

[10] Lapuschkin, S.; Binder A.; Müller, K.; Samek W.: *Understanding and Comparing Deep Neural Networks for Age and Gender Classification*, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

Lösungsansatz: vorhandenes Wissen nutzten

- Forschung über geschlechterspezifische Sprache existiert bereits.
- Könnte Algorithmen entwickeln, die geschlechterspezifische Sprache erkennt und diese bekämpft.
  - Sortierungen nach Geschlecht in Aufzählungen aufbrechen.
  - Adjektive aus Trainingsdaten entfernen.
  - "Miss" und "Mrs" könnten durch "Ms" ersetzt werden.
  - Präsenz von Frauen in Texten erhöhen.

[8] Leavy, S.: *Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning*, 28.05.2018.

Lösungen

- Ideale Datensätze erstellen
- Diversität in Unternehmen
- Forschen, um Arbeitsweise von künstlichen Neuronalen Netzen weiter zu verstehen

Zusammenfassung

- Algorithmen sind nicht objektiv, sondern reproduzieren gelernten Bias.
- Dies betrifft einen Großteil der Algorithmen, die uns umgeben.
- man muss Algorithmen kritisch hinterfragen.

Fragen an die Gesellschaft

Sollen Algorithmen über Menschen urteilen und diese bewerten können?

Wo sollten Algorithmen eingesetzt werden und wofür nicht?

Können Algorithmen als objektive Instanz genutzt werden?

Wie kann Gesellschaft Personen, die Diskriminierung durch Technik erfahren haben, unterstützen?

Quellen

(1) Angwin, J.; Larson J.; Mattu S.; Kirchner L.: *Machine Bias: There´s software used across the country to predict duture criminals. And it´s biased against black.*, in ProPublica, 23.04.2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (23.11.2020).

(2) Rekabsaz, N.; Schedl M.: *Do Neural Ranking Models Intensify Gender Bias?*, In Proceeddings of 43rd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), 15.06.2020, https://doi.org/10.1145/3397271.3401280

(3) Evers, A.: *Sexistische Algorithmen: Apple beanchteiligt weibliche Kunden.* in eRecht24, 17.11.2019, https://www.e-recht24.de/news/sonstige/11741-sexismus-apple.html, (25.11.2020).

(4) Redaktion: *Sexistische Algorithmen: Amazons künstliche Intelligenz zur Bewerberauswahl benachteiligte systematisch Frauen.* in Meedia, 17.10.2018 https://meedia.de/2018/10/17/sexistische-algorithmen-amazons-kuenstliche-intelligenz-zur-bewerber-auswahl-benachteilige-systematisch-frauen/ (25.11.2020)

(5) Rentzsch, F.: *Computer-Panne oder Diskriminierung? Warum das Passfoto dieses Asiaten abgelehnt wurde*, in Buisness Insider, 9.12.2016, https://www.businessinsider.de/politik/das-passfoto-dieses-asiaten-wurde-abgelehnt-seine-augen-seien-angeblich-geschlossen-2016-12/ (25.11.2020).

(6) Buolamwini, J. A.: *Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers*, Thesis: S.M., Massachusetts Institute of Technology, School of Architecture and Planning, 10.09.2017, http://hdl.handle.net/1721.1/114068.

Quellen

(7) Wolfgangel, E.: *Programmierter Rassismus: Woher Algorithmen ihre Vorurteile haben – und warum die so gefährlich sind.* in Zeit Online, (19.06.2018), https://www.zeit.de/digital/internet/2018-05/algorithmen-rassismus-diskriminierung-daten-vorurteile-alltagsrassismus? (24.11.2020).

(8) Leavy, S.: *Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning*, 28.05.2018, 10.1145/3195570.3195580.

(9) Crawford, K.: *Artificial Intelligence´s White Guy Problem*, in New York Times, 25.06.2018, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html (25.11.2020).

(10) Lapuschkin, S.; Binder A.; Müller, K.; Samek W.: *Understanding and Comparing Deep Neural Networks for Age and Gender Classification*, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1629-1638.

Bildquellen

(1) Folie 3: https://www.evg-online.org/meldungen/details/news/nein-zu-rassismus-und-nein-zu-gewalt-7881/

(2) Folie 4: Screenshot von https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html

(3) Folie 5: Screenshot von Google Translate https://www.google.com/search?q=googel+translate (24.11.2020, 14.00)

(4) Folie 7: Rekabsaz, N.; Schedl M.: *Do Neural Ranking Models Intensify Gender Bias?*, In Proceeddings of 43rd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), 15.06.2020, https://doi.org/10.1145/3397271.3401280

(5) Folie 9: Screenshot/Facebook/Richard Lee https://www.businessinsider.de/politik/das-passfoto-dieses-asiaten-wurde-abgelehnt-seine-augen-seien-angeblich-geschlossen-2016-12/

(6) Folie 13: Screenshot von https://www.versus.africa/post/how-data-gaps-affect-global-businesses-in-africa-part-1-of-series

(7) Folie 14: Bild von Free-Photos auf Pixabay. https://pixabay.com/de/photos/arbeitsplatz-team-gesch

(8) Folie 17: Lapuschkin, S.; Binder A.; Müller, K.; Samek W.: *Understanding and Comparing Deep Neural Networks for Age and Gender Classification*, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1629-1638.