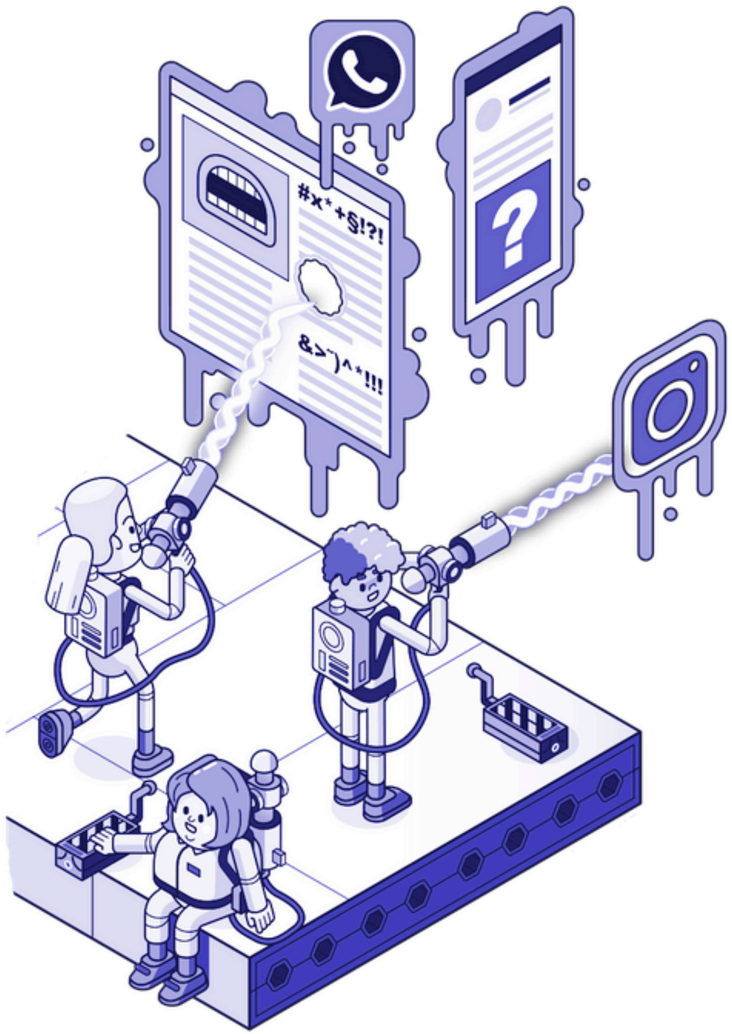
Warum automatisierte Filter rassistisch sind

Johannes Filter



Sections | The Washington Post — Democracy Dies in Darkness | Get 1 year for $29 | Sign in

The Switch

AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how.

Rep. Adam Kinzinger (R-Ill.) showed that his photo and name were used in a fake Facebook account. Zuckerberg thinks A.I. can prevent that in the future. (The Washington Post)

South China Morning Post | SIGN IN/UP

World / United States & Canada

Mark Zuckerberg believes AI will fix Facebook, but he can't say how

Mark Zuckerberg said that over the next five to 10 years, artificial intelligence would prove a champion for the world's largest social network in resolving its most pressing crises on a global scale

The Washington Post
Published: 11:15am, 12 Apr, 2018

Why you can trust SCMP

Life-size cut-outs depicting Facebook CEO Mark Zuckerberg wearing 'Fix Fakebook' T-shirts are displayed by advocacy group, Avaaz, on the southeast Lawn on Capitol Hill in Washington.

http://www.thecleaners-film.de/

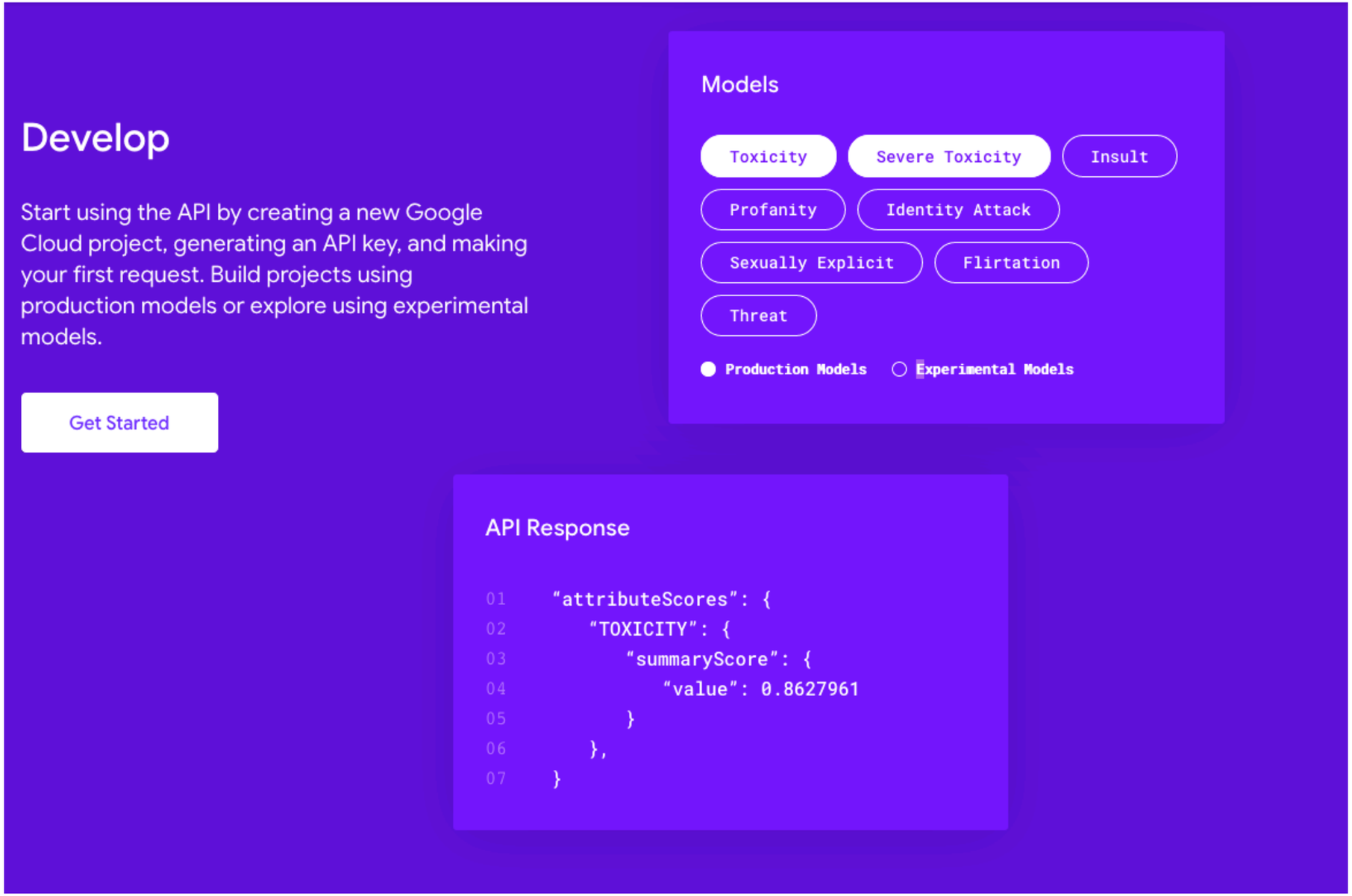
Perspective API

Perspective is an API that makes it easier to host better conversations. The API uses machine learning models to score the perceived impact a comment might have on a conversation. Developers and publishers can use this score to give realtime feedback to commenters or help moderators do their job, or allow readers to more easily find relevant information, as illustrated in two experiments below. Our first model identifies whether a comment could be perceived as “toxic” to a discussion.

Trusted for improving conversations

The New York Times DISQUS EL PAÍS Coral BY VOX MEDIA

https://www.perspectiveapi.com/



https://www.perspectiveapi.com/

17% likely to be perceived as "toxic"

all lives matter

52% likely to be perceived as "toxic"

black lives matter

https://twitter.com/danluu/status/896177697285603329

Powerful adjectives

Toxicity score given by the Perspective API to select phrases.

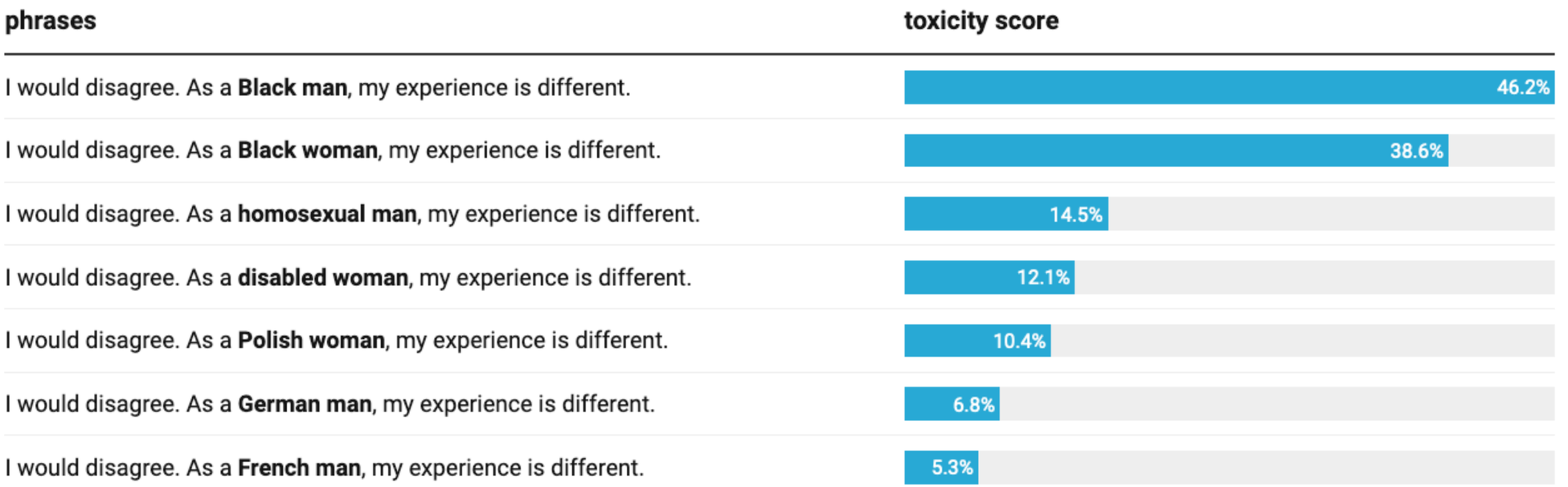
phrases
I would disagree. As a **Black man**, my experience is different.
I would disagree. As a **Black woman**, my experience is different.
I would disagree. As a **homosexual man**, my experience is different.
I would disagree. As a **disabled woman**, my experience is different.
I would disagree. As a **Polish woman**, my experience is different.
I would disagree. As a **German man**, my experience is different.
I would disagree. As a **French man**, my experience is different.

Source: AlgorithmWatch • Get the data • Created with Datawrapper

https://algorithmwatch.org/en/story/automated-moderation-perspective-bias/

Powerful adjectives

Toxicity score given by the Perspective API to select phrases.

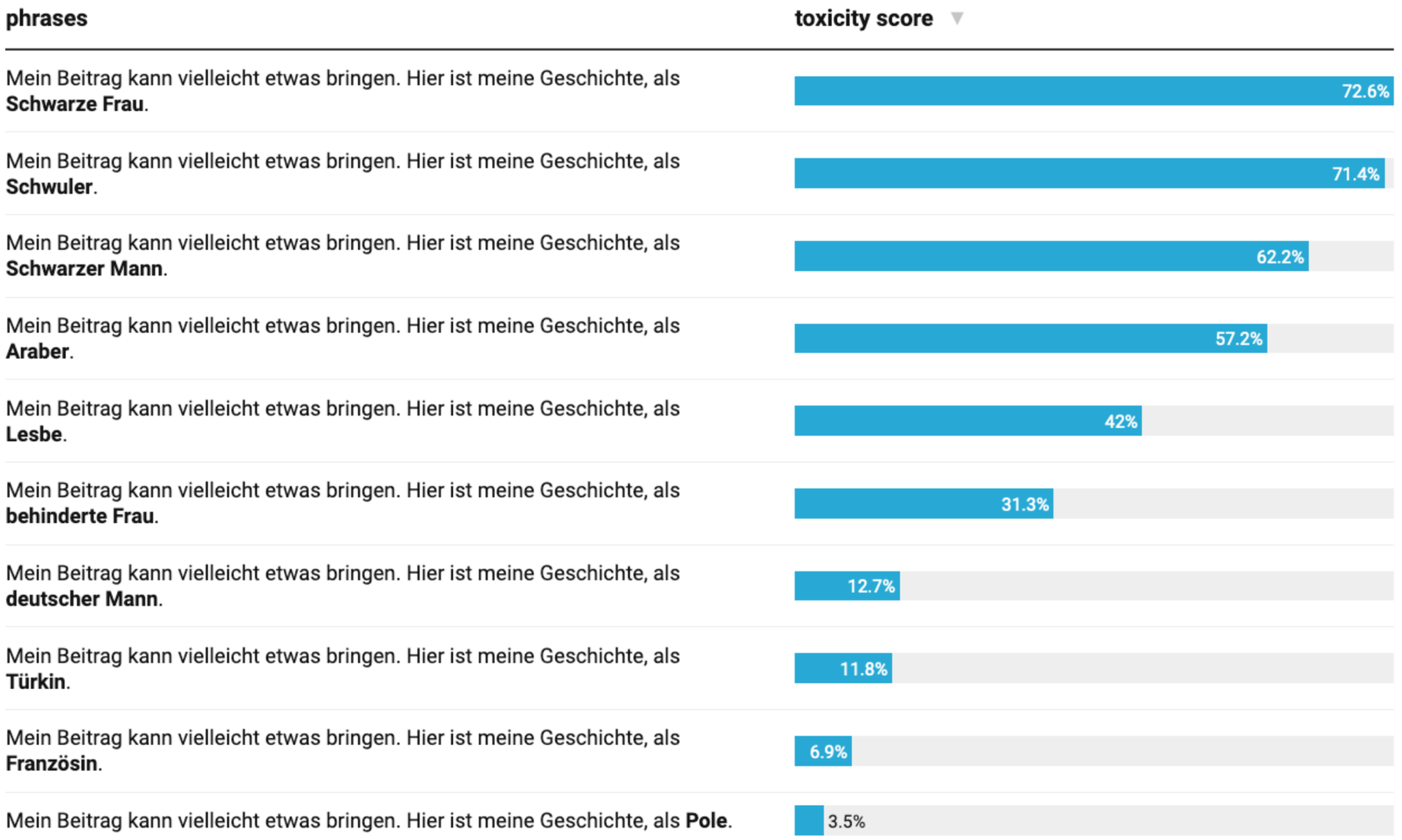


Source: AlgorithmWatch • Get the data • Created with Datawrapper

https://algorithmwatch.org/en/story/automated-moderation-perspective-bias/

German problems

Toxicity score given by the Perspective API to select phrases.



https://algorithmwatch.org/en/story/automated-moderation-perspective-bias/

NLP

Natural Language Processing

Computer-Linguistik

Stichworte

- Dummkopf
- Armleuchter

Stichworte

- ~~Dummkopf~~ Dummk0pf
- ~~Armleuchter~~ Armloichter

Verstehen

Menschliche Entscheidungen imitieren

Machine Learning / AI

Kontext

“You shall know a word by the company it keeps.”

John Rupert Firth (1957)

Beispiel

Ich esse gerne **Fisch**.

Ich esse gerne **Bananen.**

Ich esse gerne **Brot**.

Text-Corpus

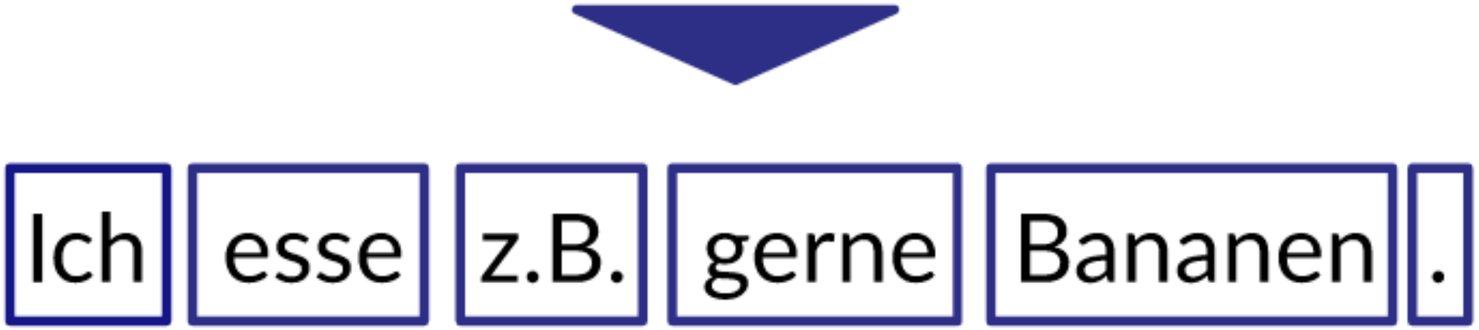
Wikipedia

Gmail

New York Times Comments

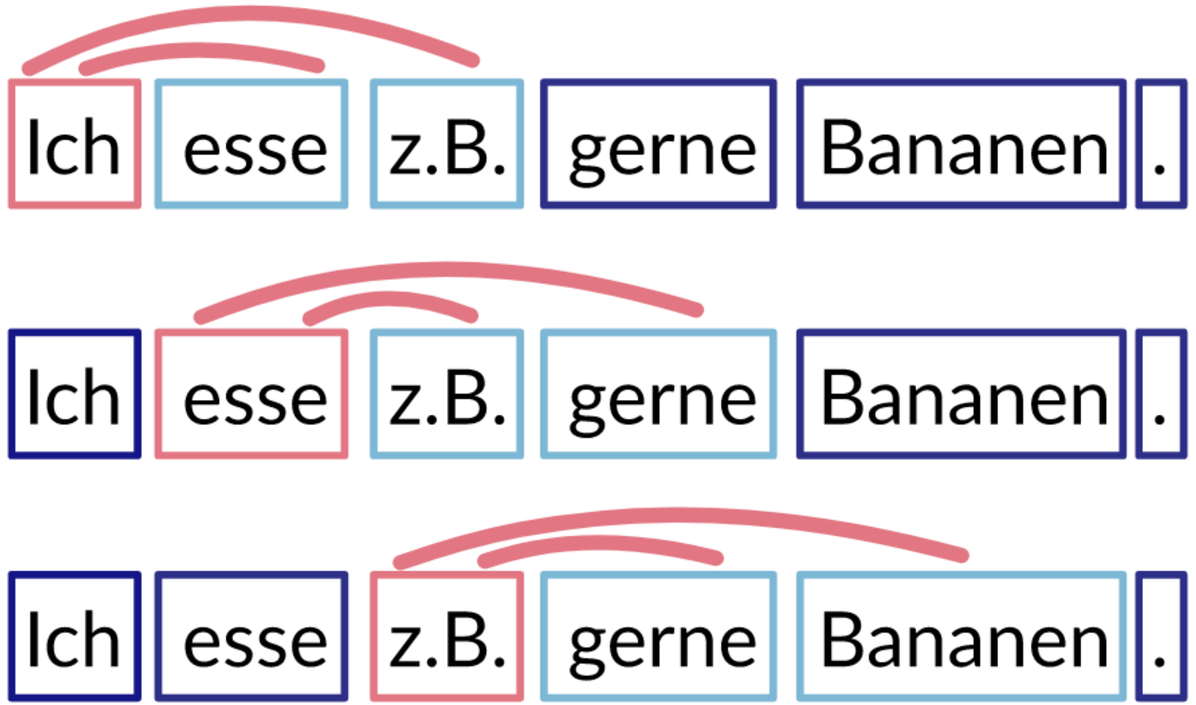
Wörter

Ich esse z.B. gerne Bananen.



Text wird durch einen Tokenizer in Wörter aufgeteilt

Zählungen



Jede Wortpaarung mit einem Abstand von zwei wird gezählt

Vektoren

gerne = [-0.1, 0.76, 0.98, -0.23, 0.3,...]

Bananen = [0.8, 0.1, 0.38, -0.5, 0.03,...]

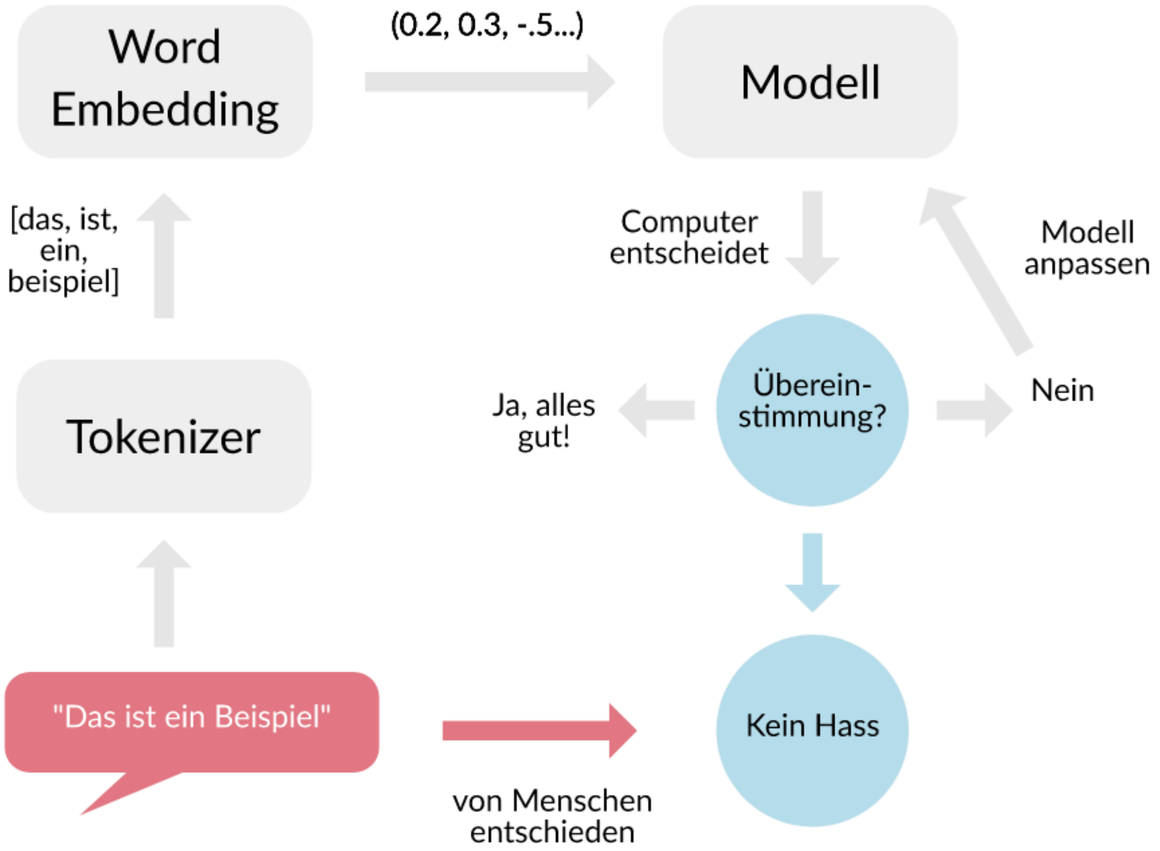
Auto = [0.12, -0.3, 0.8, 0.31, -0.9,...]

...

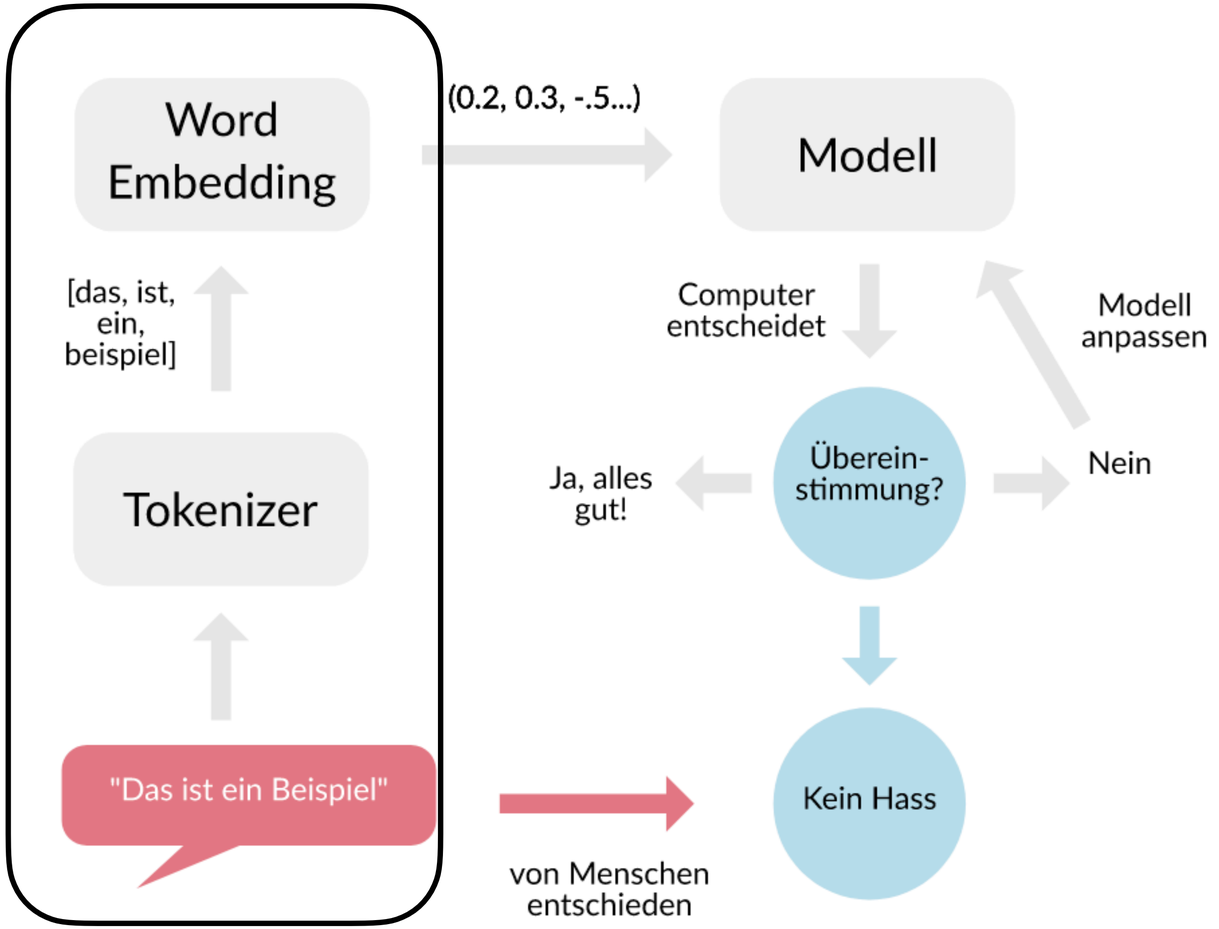
Ähnlichkeit

Die 10 ähnlichsten Wörter zu `auto` (so wie es der Computer sieht):

Wort	Ähnlichkeit
fahrzeug	84.60 %
pkw	83.74 %
fahren	72.04 %
fahrrad	71.02 %
suv	66.71 %
kfz	64.66 %
rad	63.04 %
fahrer	62.23 %
eauto	59.76 %
km	59.38 %



Der Kommentar links unten (rot) wird in das Modell gefüttert. Wenn das Modell den Kommentar so wie Menschen bewertet, passiert nichts. Mach das Modell einen Fehler, wird es angepasst.



Der Kommentar links unten (rot) wird in das Modell gefüttert. Wenn das Modell den Kommentar so wie Menschen bewertet, passiert nichts. Mach das Modell einen Fehler, wird es angepasst.

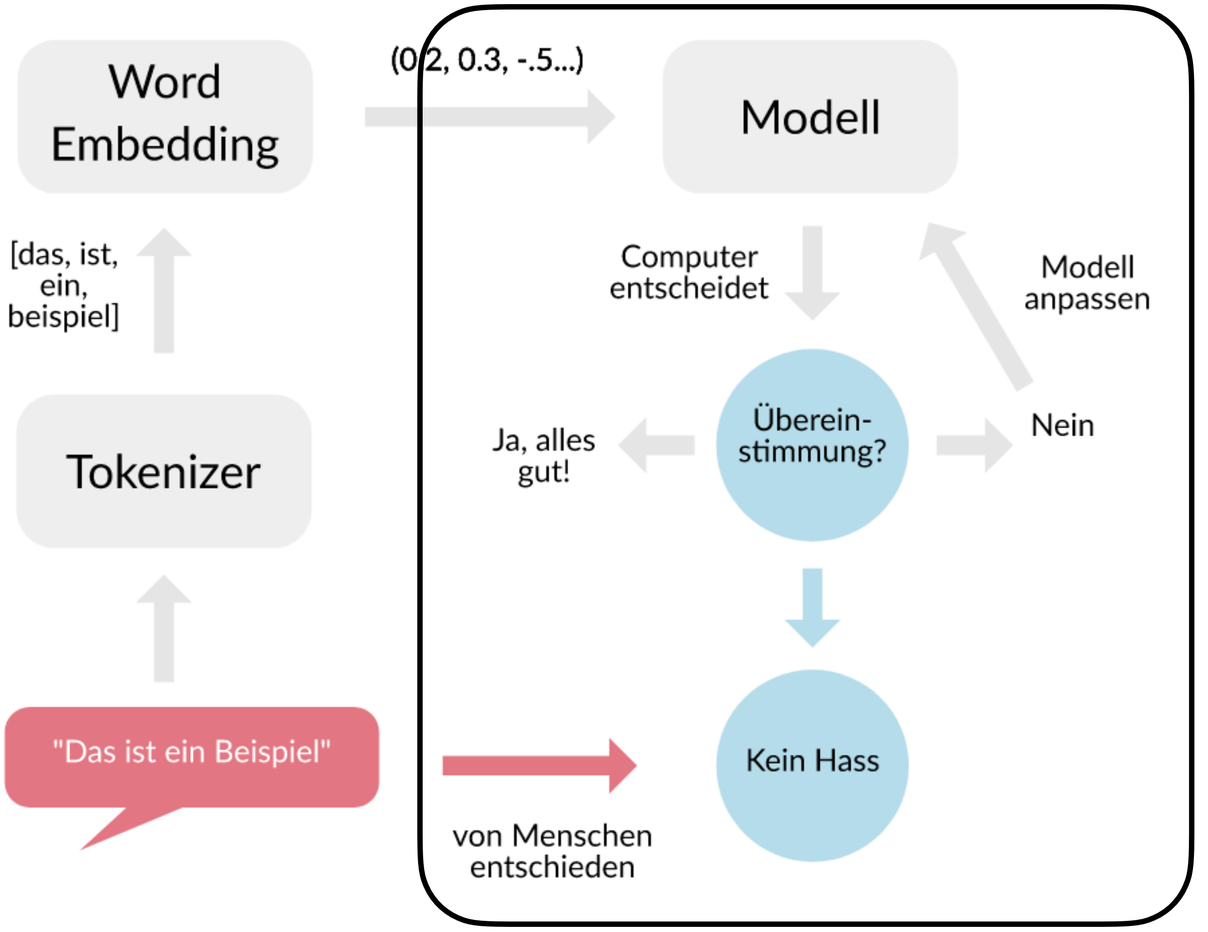
Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen[1] and **Yoav Goldberg**[1,2]
[1]Department of Computer Science, Bar-Ilan University
[2]Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com

Abstract

Word embeddings are widely used in NLP for a vast range of tasks. It was shown that word embeddings derived from text corpora reflect gender biases in society. This phenomenon is pervasive and consistent across different word embedding models, causing serious concern. Several recent works tackle this problem, and propose methods for significantly reducing this gender bias in word embeddings, demonstrating convincing results. However, we argue that this removal is superficial. While the bias is indeed substantially reduced according to the provided bias definition, the actual effect is mostly hiding the bias, not removing it. The gender bias information is still reflected in the distances between "gender-neutralized" words in the debiswer the analogy "man is to computer programmer as woman is to x" with "x = homemaker". Caliskan et al. (2017) further demonstrate association between female/male names and groups of words stereotypically assigned to females/males (e.g. arts vs. science). In addition, they demonstrate that word embeddings reflect actual gender gaps in reality by showing the correlation between the gender association of occupation words and labor-force participation data.

Recently, some work has been done to reduce the gender bias in word embeddings, both as a post-processing step (Bolukbasi et al., 2016b) and as part of the training procedure (Zhao et al., 2018). Both works substantially reduce the bias with respect to the same definition: the projection



Der Kommentar links unten (rot) wird in das Modell gefüttert. Wenn das Modell den Kommentar so wie Menschen bewertet, passiert nichts. Mach das Modell einen Fehler, wird es angepasst.

Rassismus

Automatisierte Filter sind rassistisch, weil wir in einer rassistischen Gesellschaft leben.

Fazit

Fazit

Faires automatisiertes Filtern ist schwer

Fazit

Faires automatisiertes Filtern ist schwer

kein Filtern ist auch keine Lösung

Fazit

Faires automatisiertes Filtern ist schwer

kein Filtern ist auch keine Lösung

Filtern nur unter Aufsicht

Fazit

Faires automatisiertes Filtern ist schwer

kein Filtern ist auch keine Lösung

Filtern nur unter Aufsicht

Monitoring / Research ist essenziell

Toxicity @1

Identity groups	Subgroup AUC	BPSN AUC	BNSP AUC
black gay	0.93	**0.53**	0.99
black queer	0.98	0.89	0.96
black straight	1.00	0.99	0.92
black bisexual	0.95	0.88	0.95
black homosexual	0.86	**0.44**	0.99
black heterosexual	0.96	0.87	0.95
black cis	0.99	0.99	0.91
black trans	0.96	0.90	0.95
black nonbinary	0.99	0.97	0.94
white lesbian	0.95	0.69	0.98
white gay	0.94	**0.58**	0.99
white queer	0.98	0.93	0.95
white straight	1.00	1.00	0.90
white bisexual	0.96	0.92	0.94
white homosexual	0.87	**0.47**	0.99
white heterosexual	0.96	0.90	0.95
white cis	1.00	0.99	0.90
white trans	0.97	0.94	0.94
white nonbinary	0.99	0.99	0.92

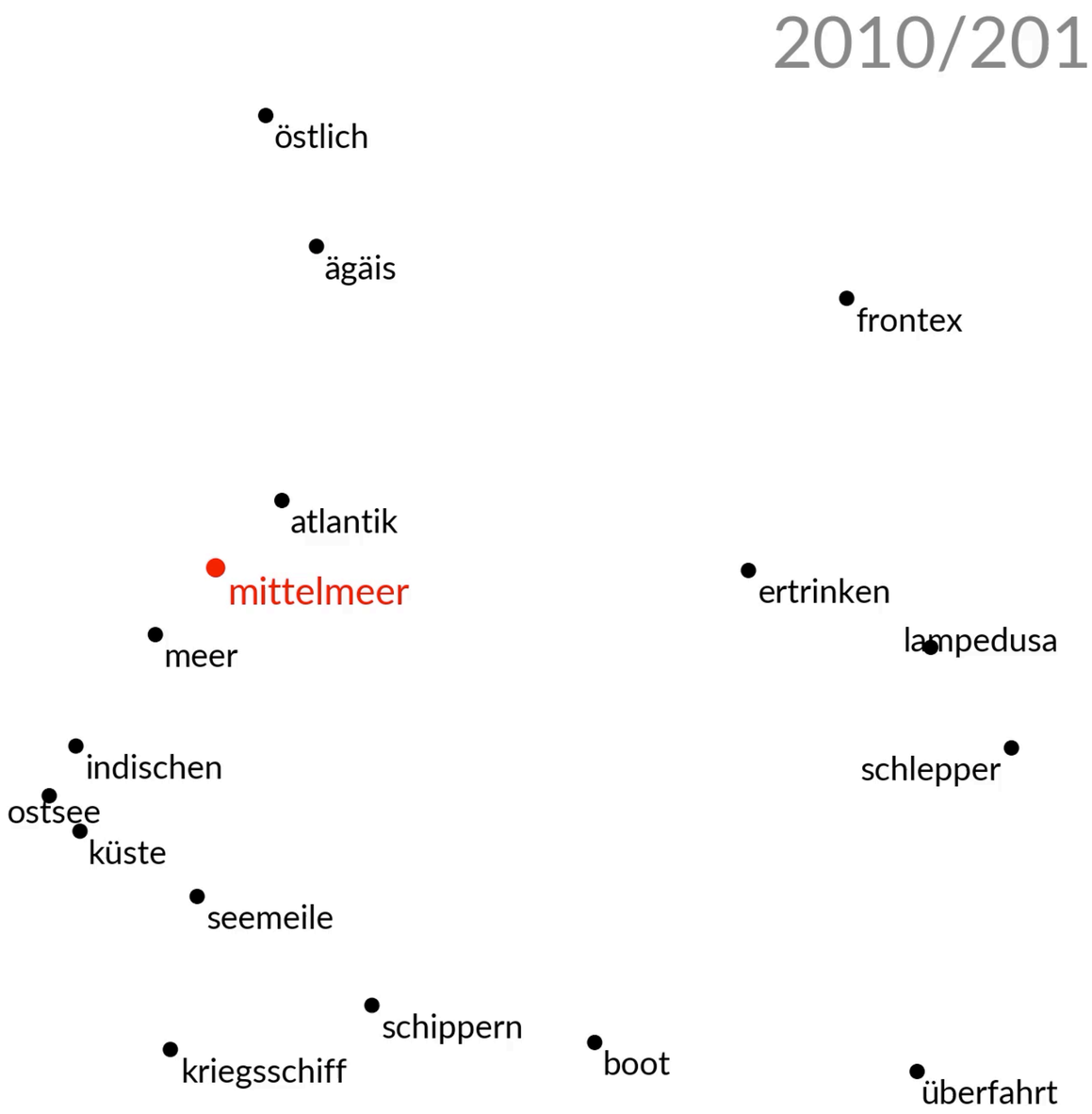
Toxicity @6

Identity groups	Subgroup AUC	BPSN AUC	BNSP AUC
black gay	1.00	0.89	1.00
black queer	0.97	0.96	0.99
black straight	0.99	0.99	0.98
black bisexual	0.95	0.93	0.99
black homosexual	1.00	0.92	1.00
black heterosexual	1.00	0.97	1.00
black cis	1.00	1.00	0.99
black trans	1.00	0.98	1.00
black nonbinary	1.00	1.00	0.99
white lesbian	1.00	0.98	1.00
white gay	1.00	0.95	1.00
white queer	1.00	0.99	0.99
white straight	1.00	1.00	0.98
white bisexual	1.00	0.98	0.99
white homosexual	1.00	0.97	1.00
white heterosexual	1.00	1.00	1.00
white cis	1.00	1.00	0.97
white trans	1.00	1.00	1.00
white nonbinary	1.00	1.00	0.98

https://support.perspectiveapi.com/s/about-the-api-model-cards

https://kommentare.vis.one

ähnliche Wörter zu »mittelmeer«



Danke für die Aufmerksamkeit!

👉 https://kommentare.vis.one

@fil_ter