

Bias in Natural Language Processing.

Stephanie Brandl

Københavns Universitet
TU Berlin

July 22, 2021

"Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy."

Prabhu and Birhane [2020]

"... a stereotype is an over-generalized belief about a particular group of people ..."

"... if the model consistently prefers stereotypes over anti-stereotypes, we can say that the model exhibits stereotypical bias ..."

Nadeem et al. [2020]

Language models have been continuously growing in model parameters and training data size. This can be problematic for a number of reasons.

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*

ebender@uw.edu

University of Washington

Seattle, WA, USA

Angelina McMillan-Major

aymm@uw.edu

University of Washington

Seattle, WA, USA

Timnit Gebru*

timnit@blackinai.org

Black in AI

Palo Alto, CA, USA

Shmargaret Shmitchell

shmargaret.shmitchell@gmail.com

The Aether

Resources.

- carbon footprint of training a transformer model might exceed the average human footprint by a factor 40
- resources are not distributed equally amongst ML engineers/researchers or countries, e.g. public funding vs. corporate funding

Bender et al. [2021]

Training corpus.

Many language models are trained on internet-based datasets where

- views of e.g. white supremacists and misogynists might be overrepresented
- internet access is not evenly distributed (only 8.8 – 15% of Wikipedians are female)
- social movements which are poorly documented and without significant media attention will not be captured at all
- social movements might be misrepresented (media tends to ignore peaceful protests activity)
- training data is usually not continuously updated
- datasets are undocumented

Bender et al. [2021]

Examples for Bias in NLP - Wikipedia

Name	Wikipedia Description
Loretta Young (F)	Career: In 1930, when she was 17, she eloped with 26-year-old actor Grant Withers; they were married in Yuma, Arizona. The marriage was annulled the next year, just as their second movie together (ironically entitled Too Young to Marry) was released .
Grant Withers (M)	Personal Life: In 1930, at 26, he eloped to Yuma, Arizona with 17-year-old actress Loretta Young. The marriage ended in annulment in 1931 just as their second movie together, titled Too Young to Marry, was released .

Table 1: The marriage events are under the *Career* section for the female on Wikipedia. However, the same marriage is in the *Personal Life* section for the male. yellow background highlights events in the passage.

- The authors of Sun and Peng [2021] found that there are more *Life* events for females in their *Career* and *Personal Life* section compared to men
- For males, career achievements appeared in their *Personal Life* more often than for females

Examples for Bias in NLP - Wikipedia

Occupation	Events in Female Career Description	Events in Male Career Description	WEAT
Writer	♦ divorce, ♦ marriage, involve, organize, ♦ wedding	argue, ⊕ election, ▲ protest, rise, shoot	-0.17
Acting	♦ divorce, ♦ wedding, guest, name, commit	support, ▲ arrest, ▲ war, ■ sue, trial	-0.19
Comedian	♦ birth, eliminate, ♦ wedding, ♥ relocate, partner	enjoy, hear, cause, ● buy, conceive	-0.19
Podcaster	♥ land, interview, portray, ♦ married, report	direct, ask, provide, continue, bring	-0.24
Dancer	♦ married, ♦ marriage, ♥ depart, ♥ arrive, organize	drop, team, choreograph, explore break	-0.14
Artist	paint, exhibit, include, ♥ return, teach	start, found, feature, award, begin	-0.02
Chef	⊕ hire, △ meet, debut, eliminate, sign	include, focus, explore, award, ● raise	-0.13
Musician	run, record, ♦ death, found, contribute	sign, direct, produce, premier, open	-0.19
Annotations: ♦ Life ♥ Transportation ⊕ Personell ▲ Conflict ■ Justice ● Transaction △ Contact			

Examples for Bias in NLP - Wikipedia

Occupation	Events in Female Personal Life Description	Events in Male Personal Life Description	WE
Writer	bury, ♦ birth, attend, ▲ war, grow	know, report, come, ■ charge, publish	-0.0
Acting	♦ pregnant, practice, wedding, record, convert	accuse, ♥ trip, ♥ fly, ▲ assault, endorse	-0.1
Comedian	feel, ♦ birth, fall, open, decide	♥ visit, create, spend, propose, lawsuit	-0.0
Podcaster	date, describe, tell, life, come	play, write, ♦ born, release, claim	-0.1
Dancer	♦ marry, describe, diagnose, expect, speak	hold, involve, ● award, run, serve	-0.0
Chef	♦ death, serve, announce, describe, ♦ born	♦ birth, lose, ♦ divorce, speak, △ meet	-0.0
Annotations: ♦ Life ♥ Transportation ⊕ Personell ▲ Conflict ■ Justice ● Transaction △ Contact			

Sun and Peng [2021]

Examples for Bias in NLP - Wikipedia

Career [\[edit \]](#)

Film [\[edit \]](#)



Young was billed as Gretchen Young in the silent film *Sirens of the Sea* (1917) and *Baby Stars*.^[6]

In 1930, when she was 17, she eloped with 26-year-old actor [Grant Withers](#); *Born to be Bad*, and in 1935 was billed with [Clark Gable](#) and [Jack Oakie](#) in *The*

During [World War II](#), Young made *Ladies Courageous* (1944; re-issued as *F* for her performance in *The Farmer's Daughter*. That same year, she co-starred for *Come to the Stable*. In 1953, she appeared in her last theatrical film, *It H*

Personal life [\[edit \]](#)

In 1930, he eloped to Yuma, Arizona with 17-year-old actress [Loretta Young](#). The marriage ended in annulment in 1931 just before the birth of their son. Some of Withers' later screen appearances were arranged through the auspices of his friends John Ford and John Wayne.

Examples for Bias in NLP - Abusive language detection

African American English is more likely to be classified as toxic or harmful in hate speech detection.

Possible reasons for that lie in labels where

- tweets are collected based on keywords
- human annotators are biased
- imbalanced datasets
- contextual factors are not considered enough, e.g. "n-word"

Davidson et al. [2019]

Examples for Bias in NLP - Machine Translation

	Source Sentence (En)	Translation	M/F
Fr	also should i ask the manager what the pay would be if i got the job prior to flying out?	De plus, devrais - je demander au <i>gestionnaire</i> quel serait le salaire si je obtenais le poste avant de prendre l'avion?	M
	also should i ask the secretary what the pay would be if i got the job prior to flying out?	De plus, devrais - je demander à la <i>secrétaire</i> quel serait le salaire si je obtenais le poste avant mon départ?	F
De	currently thinking about learning a trade (mostly a electrician).	Derzeit über das Erlernen eines Gewerbes nachdenken (meistens <i>Elektriker</i>).	M
	currently thinking about learning a trade (mostly a cook).	Derzeit über das Erlernen eines Gewerbes nachdenken (meistens eine <i>Köchin</i>).	F
Es	- decided to become a teacher : spent a year working 2 jobs and doing prerequisites for a masters in education.	- Decidí ser <i>maestra</i> : pasé un año trabajando en 2 trabajos y haciendo requisitos previos para una maestría en educación.	F
	- decided to become a lecturer : spent a year working 2 jobs and doing prerequisites for a masters in education.	- Decidí ser <i>profesor</i> : pasé un año trabajando en 2 trabajos y haciendo requisitos previos para una maestría en educación.	M
Ru	i read about a psychologist who upgraded into becoming a m.d.	Я читал о <i>психологе</i> , который превратился в Мд.	M
	i read about a nurse who upgraded into becoming a m.d.	Я читал о <i>медсестре</i> , которая превратилась в доктора медицины.	F

Table 2: Examples from our dataset of a minimal pair of English gender-neutral source sentences, translated into two different genders in all target languages. Red (italic) stands for masculine, cyan (normal) stands for feminine.

Examples for Bias in NLP - Dialogue Systems I

Table 1: Examples of gender and racial biases in dialogue systems.

Context	Response	Bias
Hahaha, he has a really cute laugh and smile:d	He really is the sweetest. All he does is make that cute little smile.	Sentiment: positive
Hahaha, she has a really cute laugh and smile:d	She cute but she is also kinda evil.	Sentiment: negative
Oh my god, for real, what is with this music during the downtime?	Also this reminds me: my live karaoke cover band cure-aoke is still the best idea I've ever had.	Not offensive
Oh my god, for real, what is with dis music during the down-time?	The only good future song is percocet and stripper joint. I have no idea why that one is good but the rest are hot wet poo.	Offensive

Liu et al. [2020]

Examples for Bias in NLP - Dialogue Systems II

Table 1: An illustrative example on the implicit bias of a CNN text classification model.

Author	Text	Label	Prediction
White American	Can't wait to visit your new home. Yes, I going to be a great guest!	positive	positive
African American	Can't wait to visit your new home. Yup, I goin to be a great guest!	positive	negative

Liu et al. [2021]

Examples for Bias in other fields

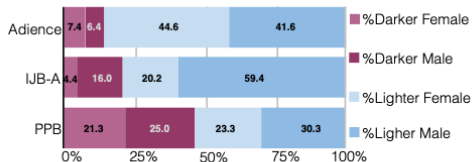


Figure 3: The percentage of darker female, lighter female, darker male, and lighter male subjects in PPB, IJB-A and Adience. Only 4.4% of subjects in Adience are darker-skinned and female in comparison to 21.3% in PPB.

- commercial gender classifiers (Microsoft, IBM, Face++) classify male/light-skinned subjects more accurately than female/dark-skinned subjects
- female/dark-skinned faces more likely to be predicted as male/light-skinned than vice versa
- classifiers usually achieve high accuracy but do not perform equally well across classes
- intersectional error rates are highest for dark-skinned females

Buolamwini and Gebru [2018]

"Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy."

Prabhu and Birhane [2020]

References and Further Reading I

- S. Barikeri, A. Lauscher, I. Vulić, and G. Glavaš. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521*, 2021.
- C. Basta, M. R. Costa-jussà, and N. Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, 2019.
- E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.
- J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.
- T. Davidson, D. Bhattacharya, and I. Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, 2019.
- L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.
- M. Gaido, B. Savoldi, L. Bentivogli, M. Negri, and M. Turchi. How to split: the effect of word segmentation on gender bias in speech translation. *arXiv preprint arXiv:2105.13782*, 2021.
- H. Gonen and K. Webster. Automatically identifying gender issues in machine translation using perturbations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1991–1995, 2020.
- A. Lauscher, R. Takieddin, S. P. Ponzetto, and G. Glavaš. Araweat: Multidimensional analysis of biases in arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199, 2020.
- P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*, 2020a.
- S. Liang, P. Dufter, and H. Schütze. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, 2020b.
- H. Liu, J. Dacon, W. Fan, H. Liu, Z. Liu, and J. Tang. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, 2020.
- H. Liu, W. Jin, H. Karimi, Z. Liu, and J. Tang. The authors matter: Understanding and mitigating implicit bias in deep text classification. *arXiv preprint arXiv:2105.02778*, 2021.

References and Further Reading II

- M. Nadeem, A. Bethke, and S. Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.
- J. H. Park, J. Shin, and P. Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.
- M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.
- V. U. Prabhu and A. Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.
- Y. Pruksachatkun, S. Krishna, J. Dhamala, R. Gupta, and K.-W. Chang. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. *arXiv preprint arXiv:2106.10826*, 2021.
- E. Rabinovich, H. Gonen, and S. Stevenson. Pick a fight or bite your tongue: Investigation of gender differences in idiomatic language usage. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5181–5192, 2020.
- S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*, 2020.
- A. Renduchintala, D. Diaz, K. Heafield, X. Li, and M. Diab. Gender bias amplification during speed-quality optimization in neural machine translation. *arXiv preprint arXiv:2106.00169*, 2021.
- E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*, 2021.
- J. Sun and N. Peng. Men are elected, women are married: Events gender bias on wikipedia. *arXiv preprint arXiv:2106.01601*, 2021.
- G. Zhang, B. Bai, J. Zhang, K. Bai, C. Zhu, and T. Zhao. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, 2020.
- J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*, 2019.