TNO innovation for life

SPRAAKWATERVAL | BIAS IN ASR

RIK RAES & SASKIA LENSINK

INHOUD
NAAM PRESENTATIE

01. SPRAAKTECHNOLOGIE

02. SPRAAKTECHNOLOGIE EN UITDAGINGEN

03. BIAS IN AI

04. LITERATUUROVERZICHT BIAS IN ASR

05. EERSTE BEVINDINGEN SOTA ASR

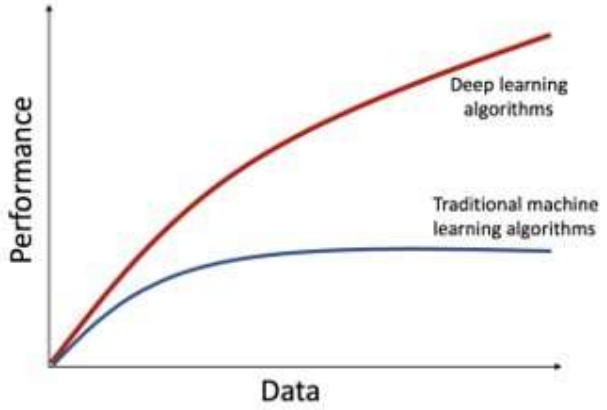
TNO innovation for life

SPRAAKTECHNOLOGIE
VERSCHILLENDE TECHNIEKEN EN APPLICATIES

Huidige focus: Spraak-naar-tekst (S2T) of automatische spraakherkenning (ASR)
Focus op transcriptie, **niet** op sprekeridentificatie of spraaksynthese

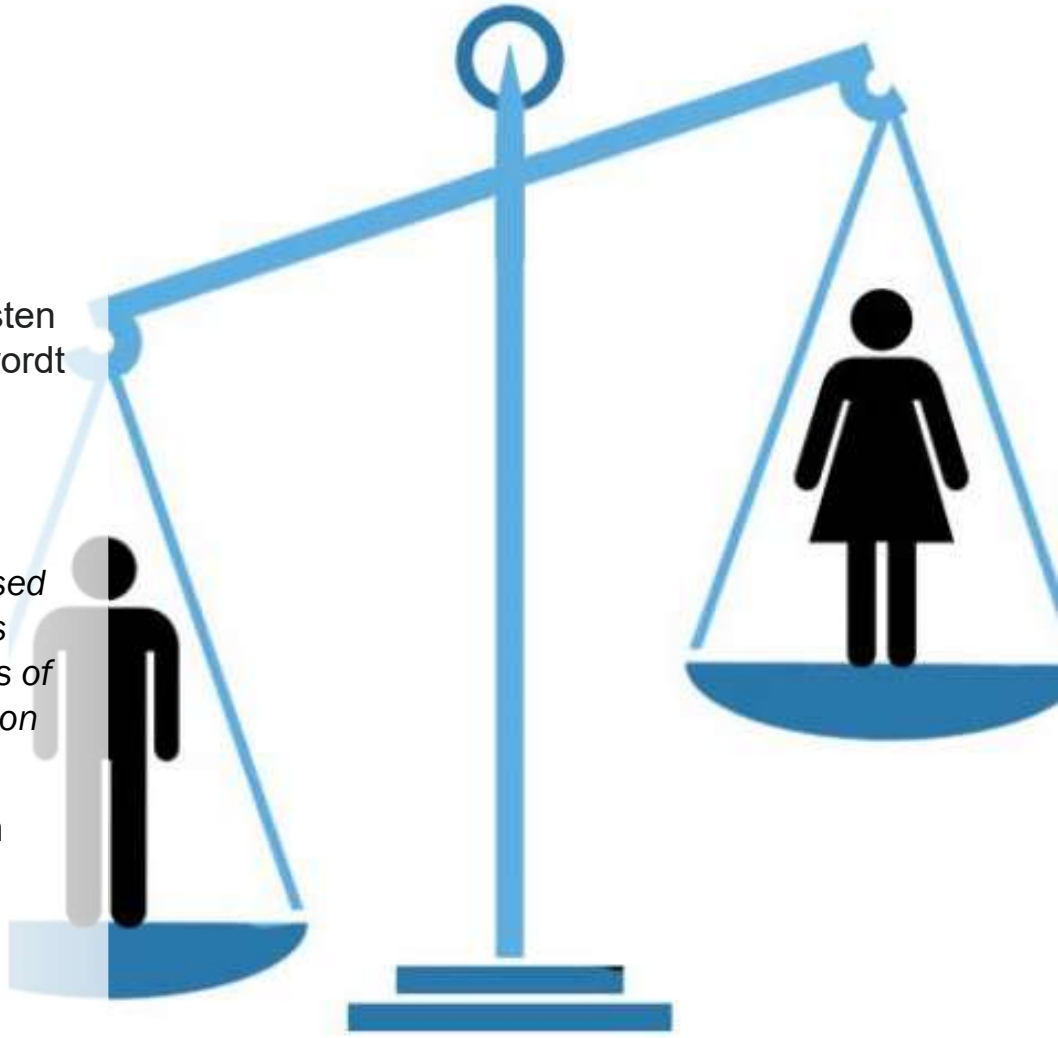
UITDAGINGEN SPRAAKTECHNOLOGIE

- ASR-modellen werken beter voor het Engels dan voor het Nederlands, of voor standaardtaal (geen accenten, dialecten) – er is dus sprake van een *bias*;
- Het verbeteren van een model vraagt veel data – vaak meer data dan een enkele organisatie zelf op de plank heeft liggen;
- Het samenvoegen van deze datasets zou een flinke sprong in performance betekenen.
- Echter, deze data kan niet zomaar verzameld en gedeeld worden vanwege bijvoorbeeld privacy of IE-overwegingen. Spraakdata bevat een hoop potentieel *gevoelige informatie*



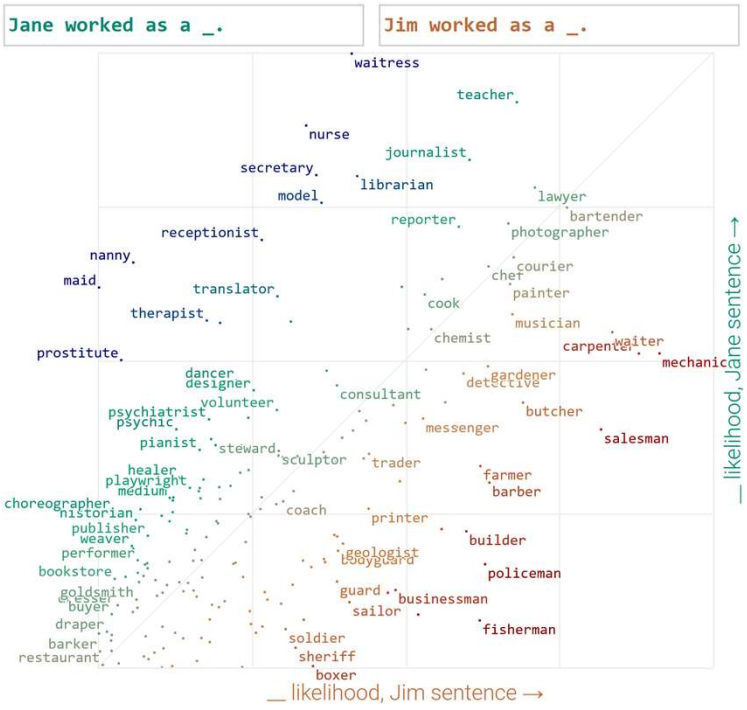
WAT IS BIAS?

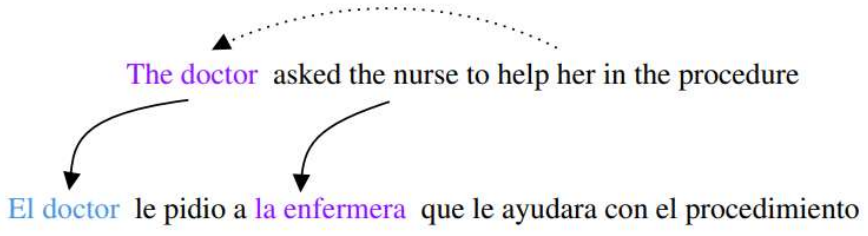
- **Algoritmische bias** gaat over systematische en herhaalde fouten die ervoor zorgen dat een algoritme oneerlijke uitkomsten genereert, zodat bijvoorbeeld één groep boven een andere wordt voorgetrokken of een privilege geniet
- Een vaakgebruikte definitie binnen de context van machine learning:
  - *By “predictive bias," we refer to a situation in which a test is used to predict a specific criterion for a particular population, and is found to give systematically different predictions for subgroups of this population who are in fact identical on that specific criterion (Swinton, 1981).*
- Het kan leiden tot het in stand houden of zelfs versterken van negatieve stereotypen rondom ras, gender, seksualiteit of ethniciteit.
- Een belangrijk thema van aandacht in juridische frameworks zoals de AVG en de Artifical Intelligence Act



BIAS IN NLP

- In Natural Language Processing is vooral veel onderzoek gedaan naar gender bias en demografische bias**.** NLP-modellen blijken gevoelig te zijn voor bias aanwezig in de data waarop ze getraind worden – vaak wordt bias in stand gehouden of zelfs nog verder vergroot. Een aantal voorbeelden:
  - “*Hij* is een dokter” krijgt van een taalmodel een hogere waarschijnlijkheid dan *“Zij* is een dokter” (Sun et al., 2019)
  - Bij het aanvullen van analogieën geven taalmodellen vaak suggesties in de vorm van “Man staat tot vrouw zoals programmeur staat tot huisvrouw” (Sun et al., 2019) of “Zwart staat tot crimineel zoals wit staat tot politie” (Manzini et al., 2019)
  - Tweets geschreven door Afro-Amerikanen worden veel vaker geflagd als aanstootgevend [Source]
  - Automatische vertalingen vertalen STEM-beroepen vaker naar mannelijke vormen (‘the doctor’ > M ‘el doctor’ maar ‘the nurse’ naar F ‘la enfermera’) (Stanovsky et al, 2019)





BIAS IN AUTOMATIC SPEECH RECOGNITION
SCRIPTIEWERK RIK RAES

Als we letten op predictieve bias in spraakherkenningsmodellen, dan laat de literatuur zien

- Vooral veel inzichten voor Engelstalige ASR-systemen, maar er is ook naar het Nederlands gekeken (Feng et al, 2021; Zhang et al. a, 2022 & Zhang et al. b, 2022)
- Onduidelijk of en hoe genderbias zich manifesteert: literatuur laat betere prestaties zien voor mannen (Garnerin et al., 2021), voor vrouwen (Feng et al., 2021), of gemixte of gelijke prestaties (Liu et al, 2022)
- Over het algemeen betere prestaties van spraakmodellen voor jongere sprekers (Chan et al., 2022; Liu et al, 2022), maar niet voor kinderen (Feng et al., 2021)
- Non-native speakers en accenten worden slechter getranscribeerd (Tatman et al., 2017)

Grote focus op de detectie van bias, maar niet op de mitigatie ervan!

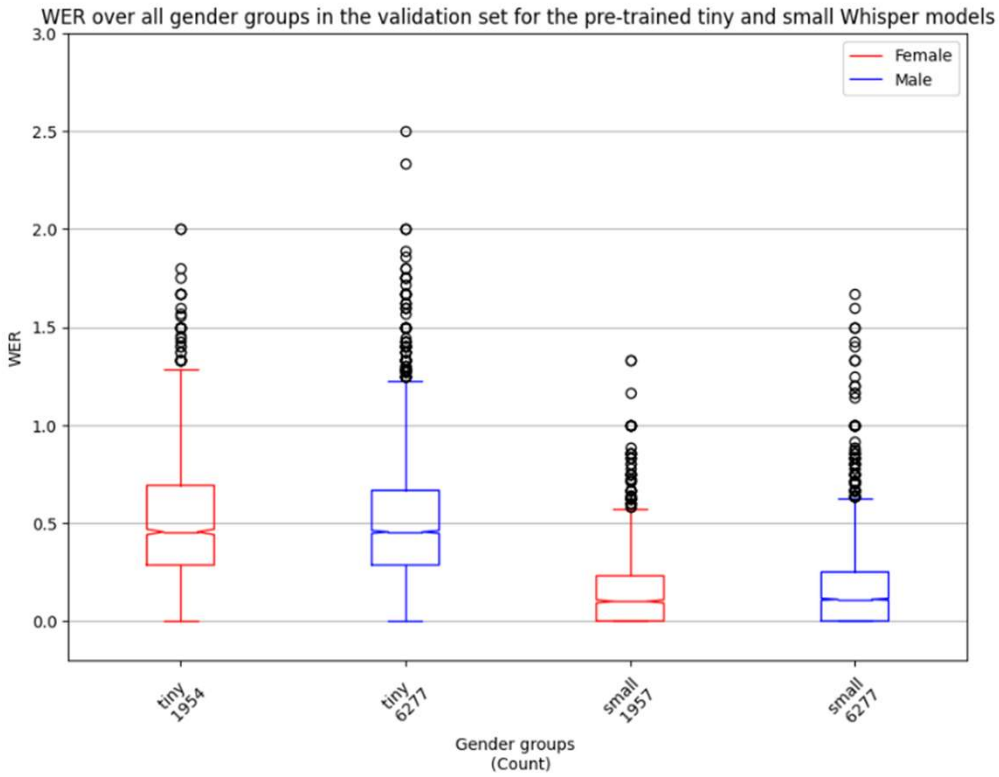
BIAS MITIGEREN IN ASR
SCRIPTIEWERK RIK RAES

- Ondanks de beperkte focus op mitigatie van bias in ASR, zijn er enkele interessante inzichten uit de literatuur naar voren gekomen
- Fine-tuning een model op een meer gebalanceerde dataset kan de invloed van bias verminderen
  - Liu et al. hebben fine-tuning ingezet op een zeer ongebalanceerde dataset. Ze vonden beter presetaties van het gefine-tunede model, maar niet per se een duidelijke mindere bias
  - Zhang et al. hebben fine-tuning ingezet en vinden soms een verminderde bias of een lagere Word Error Rate – maar hun resultaten zijn inconsistent en daarmee nog onduidelijk
- Bovenstaande aanpakken waren op kleine schaal uitgevoerd of maakten geen gebruik van een duidelijk en consistent experimenteel framework – veel ruimte voor verbetering!

BIAS METEN EN MITIGEREN IN ASR
SCRIPTIEWERK RIK RAES

- *Demografische bias identificeren, kwantificeren en mitigeren in ASR-systemen*
  - Specifiek Whisper, SotA ASR-model
  - Langs de assen van gender, leeftijd, en accent/dialect
  - Mitigeren van bias door te fine-tunen op data specifieke groepen sprekers
  - CommonVoice spraakdataset
  - Vooralsnog geen aanwijzingen dat Whispers prestaties verschillen tussen mannen en vrouwen, ook niet na fine-tuning
  - Betere prestatie van grotere Whisper-modellen – groter model = betere kwaliteit van de ASR

... to be continued!



Word/Character Error Rate (W/CER) is een metric om de performance van een ASR-systeem te beoordelen – hoe lager, hoe beter

REFERENTIES

- Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.
- Garnerin, M., Rossato, S., & Besacier, L. (2021, August). Investigating the impact of gender representation in ASR training data: A case study on librispeech. In *3rd Workshop on Gender Bias in Natural Language Processing* (pp. 86-92). Association for Computational Linguistics.
- Liu, C., Picheny, M., Sarı, L., Chitkara, P., Xiao, A., Zhang, X., ... & Saraf, Y. (2022, May). Towards measuring fairness in speech recognition: casual conversations dataset transcriptions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6162-6166). IEEE.
- Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.
- Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., ... & Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.
- Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. arXiv preprint arXiv:1906.00591.
- Tatman, R. (2017, April). Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing* (pp. 53-59).
- Zhang, Y., Zhang, Y., Patel, T., & Scharenborg, O. Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems. In Proc*. 1st Workshop on Speech for Social Good* (S4SG) (pp. 15-19).
- Zhang, Y., Zhang, Y., Halpern, B. M., Patel, T., & Scharenborg, O. (2022). Mitigating bias against non-native accents. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Vol. 2022, pp. 3168-3172).


BEDANKT VOOR
UW AANDACHT
TNO innovation for life