Literature List PS/HS Bias (WS 2019/20)

Katja Markert

October 28, 2019

29.10.2019: Topics, Administration

5.11.2019 Bias and Fairness Definitions, Measuring Bias, Legal Background

Obligatory Read: (Obermeyer et al., 2019) and the representation and forum discussion in the German magazine "Der Spiegel" under `https://www.spiegel.de/netzwelt/apps/usa-algorithmus-benachteiligt-afroamerikanische-patienten-a-1293382.html`

Some example cases of societal impact of bias in algorithms: (Obermeyer et al., 2019; Klare et al., 2012; Kay et al., 2015; Buolamwini & Gebru, 2018)

Literature for legal background (US): (Barocas & Selbst, 2016; House, 2014; 2016)

Link to the German "Allgemeine Gleichstellungsgesetz": `https://www.antidiskriminierungsstelle.de/SharedDocs/Downloads/DE/publikationen/AGG/agg_gleichbehandlungsgesetz.pdf?__blob=publicationFile`

Survey paper on bias and fairness in machine learning: (Mehrabi et al., 2019)

Overview paper on gender bias in NLP systems: (Sun et al., 2019)

The industry perspective: (Holstein et al., 2019)

12.11.2019: Measuring Bias in Word Embeddings (Basis)

Literature: (Garg et al., 2018; Caliskan et al., 2017)

19.11.2019 and 26.11.2019: Mitigating Bias in Word Embeddings

Mitigation via in- or post-processing: (Bolukbasi et al., 2016; Zhao et al., 2018b; 2019; Manzini et al., 2019)

Mitigation via data modification: (Brunet et al., 2019; Maudslay et al., 2019)

Does it really work?: (Gonen & Goldberg, 2019)

3.12.2019: Selection Bias: Bias in Wikipedia (Basis)

Literature: (Callahan & Herring, 2011) on cultural bias in Wikipedia, (Otterbacher, 2015) on IMDB biography bias, (Bamman & Smith, 2014; Wagner et al., 2015; 2016) on gender bias in Wikipedia

10.12.2019: Evaluation Corpora: GBETS (Gender Bias Evaluation Test Sets) for Coreference Resolution

Literature: (Webster et al., 2018; Rudinger et al., 2018; Zhao et al., 2018a)

17.12.2019: Case Studies I: Bias in MT and dialect processing

Dialect leading to biased processing: (Blodgett et al., 2016; 2018; Sap et al., 2019)

Machine Translation: (Escudé Font & Costa-jussà, 2019; Prates et al., 2018; Vanmassenhove et al., 2018)

7.1.2020 and 14.1.2020 Bias as disparate impact of machine learning classification (Mostly advanced)

Literature (classic and seminal papers): (Calders & Verwer, 2010; Dwork et al., 2012; Kamishima et al., 2012; Feldman et al., 2015; Hardt et al., 2016)

Very Advanced Literature: (Elazar & Goldberg, 2018) on using adversarial training to prevent leakage of protected attributes, (Zafar et al., 2017) on formalizing and then mitigating a new notion called disparate mistreatment

20.2.2020 Case Studies II: Bias in Text Classification (Mostly Basis)

Literature: (Dixon et al., 2018; Park et al., 2018; Kiritchenko & Mohammad, 2018)

28.1.2020 Case Studies III: Visual semantic role labeling and/or image search results (Mostly advanced)

In Visual semantic role labeling: (Zhao et al., 2017)

In Image search results: (Kay et al., 2015)

4.2.2020: Final Discussion and Project Ideas

References

Bamman, David & Noah A Smith (2014). Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

Barocas, Solon & Andrew D Selbst (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671.

Blodgett, Su Lin, Lisa Green & Brendan O'Connor (2016). Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130.

Blodgett, Su Lin, Johnny Wei & Brendan OConnor (2018). Twitter universal dependency parsing for african-american and mainstream american english. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1415–1425.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama & Adam T Kalai (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357.

Brunet, Marc-Etienne, Colleen Alkalay-Houlihan, Ashton Anderson & Richard Zemel (2019). Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*, pp. 803–811.

Buolamwini, Joy & Timnit Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91.

Calders, Toon & Sicco Verwer (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292.

Caliskan, Aylin, Joanna J Bryson & Arvind Narayanan (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Callahan, Ewa S & Susan C Herring (2011). Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915.

Dixon, Lucas, John Li, Jeffrey Sorensen, Nithum Thain & Lucy Vasserman (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, ACM.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold & Richard Zemel (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, ACM.

Elazar, Yanai & Yoav Goldberg (2018). Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 11–21.

Escudé Font, Joel & Marta R Costa-jussà (2019). Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger & Suresh Venkatasubramanian (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, ACM.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky & James Zou (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Gonen, Hila & Yoav Goldberg (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 609–614.

Hardt, Moritz, Eric Price, Nati Srebro et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323.

Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik & Hanna Wallach (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, p. 600, ACM.

House, The White (2014). *Big Data: Seizing Opportunities, Preserving Values.* Technical Report: The White House.

House, The White (2016). *Big Data: A report on Algorithmic Systems, Opportunity and Civil Rights.* Technical Report: The White House.

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh & Jun Sakuma (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50, Springer.

Kay, Matthew, Cynthia Matuszek & Sean A Munson (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3819–3828, ACM.

Kiritchenko, Svetlana & Saif Mohammad (2018). Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 43–53.

Klare, Brendan F, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge & Anil K Jain (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801.

Manzini, Thomas, Lim Yao Chong, Alan W Black & Yulia Tsvetkov (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 615–621.

Maudslay, Rowan Hall, Hila Gonen, Ryan Cotterell & Simone Teufel (2019). It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of RMNLP 2019*.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman & Aram Galstyan (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Obermeyer, Ziad, Brian Powers, Christine Vogeli & Sendhil Mullainathan (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Otterbacher, Jahna (2015). Linguistic bias in collaboratively produced biographies: crowdsourcing social stereotypes? In *ICWSM*, pp. 298–307.

Park, Ji Ho, Jamin Shin & Pascale Fung (2018). Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2799–2804.

Prates, Marcelo OR, Pedro H Avelar & Luís C Lamb (2018). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, pp. 1–19.

Rudinger, Rachel, Jason Naradowsky, Brian Leonard & Benjamin Van Durme (2018). Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 8–14.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi & Noah A Smith (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678.

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang & William Yang Wang (2019). Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Vanmassenhove, Eva, Christian Hardmeier & Andy Way (2018). Getting gender right in neural machine translation. In *Proc. of EMNLP 2018*.

Wagner, Claudia, David Garcia, Mohsen Jadidi & Markus Strohmaier (2015). It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *ICWSM*, pp. 454–463.

Wagner, Claudia, Eduardo Graells-Garrido, David Garcia & Filippo Menczer (2016). Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5(1):5.

Webster, Kellie, Marta Recasens, Vera Axelrod & Jason Baldridge (2018). Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez & Krishna P Gummadi (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180, International World Wide Web Conferences Steering Committee.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez & Kai-Wei Chang (2019). Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 629–634.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez & Kai-Wei Chang (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez & Kai-Wei Chang (2018a). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2.

Zhao, Jieyu, Yichao Zhou, Zeyu Li, Wei Wang & Kai-Wei Chang (2018b). Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4847–4853.