Ethics in NLP: Bias and Dual Use

Michael Strube (Heidelberg Institute for Theoretical Studies) &
Shimei Pan (University of Maryland, Baltimore County)
michael.strube ät h-its.org
shimei ät umbc.edu

February 9, 2023

Contents

NLP applications are widely used in everyday life: web search, grammar correction, machine translation, chatbots/virtual assistants etc,. They are commonly available on our computers and mobile phones. Moreover, very large pretrained language models such as BERT and GPT-3 are at the core of many applications that understand and generate natural language. Since these models are mostly trained on human-generated data (e.g., text from the web/social media), they frequently inherit human biases and prejudices. In this seminar, we will discuss the implications of this. We will answer questions such as "How can we assess the bias in NLP models and data?" and " How to debias language models and NLP applications?" Bias assessment and mitigation will be the focus of the first half of the seminar. The second half will be dedicated to dual use: NLP helps not only us, but also e-commerce to get to know more about their customers, the industry to place personalized advertisements, authoritarian governments to censor posts in microblogs and social networks, secret services to search phone calls and emails not only for keywords but also for contents. In the seminar we will look at methods and applications from sentiment analysis, machine translation, text mining, NLP and social media, NLP in health applications, etc. We will question their ethical implications and their impact on society.

Literature

- Blue, Ethan et al. (2014). Engineering and War: Militarism, Ethics, Institutions, Alternatives. Morgan and Claypool Publishers.
- Church, Kenneth Ward and Kordoni, Valia (2021). Emerging trends: Ethics, intimidation, and the Cold War. In Natural Language Engineering, 27, pp.379-390.
- Noble, Safiya Umoja (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press.
- Caliskan, Aylin and Bryson, Joanna J and Narayanan, Arvind (2017). Semantics Derived Automatically from Language Corpora Contain Human-like Biases. In Science, 356, pp.183-186.

Dates, Topics

18.10.22
MS/SP: Motivation, overview, and administrative issues

25.10.22
MS: History Dual Use in NLP, 2nd World War, Vietnam, Funding, Snowdon, Meta Data, ...
SP: History of fair AI/NLP and case study (e.g., COMPAS).

to be prepared: Write your opinion on the Edward Snowdon case
or
possible biases in the Google search engine.
(at most 1 page)

01.11.22
No seminar. Allerheiligen (holiday)

08.11.22
Presenter: **Malte Schlenker** – The Ontological Interpretation of Informational Privacy (Floridi, 2005)

SP: Bias/fairness definitions and measures
(Hardt et al., 2016; Kusner et al., 2017; Garg et al., 2019; Foulds et al., 2020; Czarnowska et al., 2021)

To be prepared: (Floridi, 2005) (available on seminar webpage under "weitere Kursmaterialien")

15.11.22
SP: Bias mitigation in fair AI/NLP
Presenter: **Irina Wüst** (Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP.)
(Zhang et al., 2018; Dixon et al., 2018; Islam et al., 2021; Schick et al., 2021; Qian et al., 2021)

To be prepared: (Schick et al., 2021) (available on the ACL Anthology)
OR (Dixon et al., 2018) (available on seminar webpage under "weitere Kursmaterialien")

22.11.22
No seminar. MS @ HITS group leaders' retreat. SP @ Thanksgiving

29.11.22
SP: Assessing and mitigating bias in word embedding

Presenter: **Johannes Eschbach-Dymanus** (Learning gender-neutral word embedding)
(Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Zhao et al., 2018b; Gonen & Goldberg, 2019)

To be prepared: (Zhao et al., 2018b)
OR (Bolukbasi et al., 2016) (available on seminar webpage under "weitere Kursmaterialien")

06.12.22
MS/SP @ EMNLP

13.12.22
SP: Assessing and mitigating bias in pre-trained Language Models
Presenter: **Tai Mai**

(Huang et al., 2020; Liang et al., 2021; Ahn & Oh, 2021; Nadeem et al., 2021; Delobelle et al., 2022)

To be prepared: (Nadeem et al., 2021)
OR
(Liang et al., 2021) (available on seminar webpage under "weitere Kursmaterialien")

20.12.22

SP: Bias in common NLP tasks (e.g., POS taggers, parsers, coreference resolution, language detection)
Presenters:
Stefan Werner(Gender bias in POS Tagging and Dependency Parsing)
Hanna Dzhurynska(Gender Bias in Coreference Resolution)
(Jurgens et al., 2017a; Rudinger et al., 2018; Blodgett et al., 2018; Zhao et al., 2018a; Garimella et al., 2019)

To be prepared:
(Garimella et al., 2019)
OR
(Rudinger et al., 2018)

10.01.23
MS: Microblogs – Entities, Content
(Benton et al., 2016; Çarık & Yeniterzi, 2022)
survey on personalization in NLP: (Flek, 2020)
Presenter:
Blanca Birn(contextually personalised classification).
Andrea Hönig(A Twitter corpus for named entity recognition in Turkis)

17.01.23
MS: Microblogs: Profiling and Anti-Profiling

Presenter: **Janosch Gehring** (Jurgens et al., 2017b) (Writer Profiling Without the Writer's Text)

(Brennan et al., 2012; Afroz et al., 2014; Reddy & Knight, 2016; Jurgens et al., 2017b; Del Tredici et al., 2019; Mahmood et al., 2020; Emmery et al., 2021; Mubarak et al., 2022)

MS: Psychology and NLP – Lies, Depression, Suicide Prevention, Power

Presenter: **Anna Fischer** (Nguyen et al., 2022) (Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires)

To be prepared: (Brennan et al., 2012)
OR
(Shing et al., 2018)

24.01.23
MS @ Cambridge
SP: Hate speech (from both technical and bias aspect)
Presenters:
Melis Çelikkol
Lisa Jockwitz

(Waseem & Hovy, 2016; Waseem, 2016; Waseem et al., 2017; Sap et al., 2020; Zhou et al., 2021; Sap et al., 2019)

To be prepared:
(Waseem & Hovy, 2016): Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter
OR
(Sap et al., 2019): The risk of racial bias in hate speech detection.

31.01.23
Presenter: **Yi Wan Teh**
MS: Medicine and NLP – Anonymization and De-Anonymization (Lee et al., 2022)

Presenter: **Timothy Müller**
MS: Is it ethical to develop an ethics bot? (Jiang et al., 2022) – Response: (Talat et al., 2022)

To be prepared: (Lee et al., 2022)
OR
(Jiang et al., 2022)

07.02.23
Presenter: **Dang Nguyen**
MS: Model Card (Mitchell, Wallach), IBM FactSheets 360 (IBM Fair AI)
(Jernite et al., 2022)

To be prepared:
(Jernite et al., 2022)
OR
The background and limitations of GPT/ChatGPT:
(a) A paper on a model called InstructGPT by OpenAI (Ouyang et al., 2022). ChatGPT uses the same methods as InstructGPT, but with slight differences in the data collection setup.
OR
(b) A paper on the strength and limitations of LLMs like GPT/ChatGPT (Mahowald et al., 2023).

14.02.23
Discussion and Conclusions (the first half will be dedicated to the ChatGPT discussion)

To be prepared: Two questions on any topic discussed in the seminar. We'll try to address some of these.

Further Remarks:

Assessment:

1. For each class read the material marked in the schedule as *to be prepared*. Formulate two questions about the material and send them to us via email until the Monday before the class, 13:00 at the latest. If you present in class that day, you do not have to hand in questions. – Participate actively in the class (important!).

2. Choose a topic in the schedule you want to present in class. Select one or more papers from the reading list. Present this work in the class (30 minutes presentation, 15 minutes discussion).

3. Choose a second topic and present it in class. Or: Write a report/an essay towards the end of the term either about the topic you presented in class or about a new topic (6LP: 8-10 pages; 8LP (if possible): 12-15 pages)). Or: Perform an experiment/implement a system, evaluate it in comparison to a baseline, and write a very short report (3-4 pages). Here: Issues related to ethics/bias very important.

4. **Deadline for essay/report: March 31st, 2023**

Literature: Most papers can be downloaded from the *ACL Anthology* (`http://acl.ldc.upenn.edu/`), in particular all papers presented at *(E/NA)ACL*, *Coling* and *EMNLP* conferences, all workshops organized during these conferences and the journals *TACL* and *Computational Linguistics*. Papers published throuhg *AAAI* (AAAI conference, AAAI workshops, AAAI symposia, etc.) are availabe through the *AAAI Digital Library* (`http://www.aaai.org/Library`). ACM conference and journal papers can be found at the *ACM Digital Library* (`https://dl.acm.org/`). NeurIPS proceedings can be found here (`https://papers.nips.cc/`). – Other journals are available electronically at the university library (`https://www.ub.uni-heidelberg.de/`, `http://rzblx1.uni-regensburg.de/ezeit/search.phtml?bibid=UBHE&colors=3&lang=de`).
Please contact the professors if you can not find a particular paper online or at the university library.

Office hours: Right after class, or in our offices at HITS (`https://www.h-its.org/en/`).

References

Afroz, Sadia, Aylin Islam Caliskan, Ariel Stolerman, Rachel Greenstadt & Damon McCoy (2014). Doppelgänger finder: Taking stylometry to the underground. In *2014 IEEE Symposium on Security and Privacy, SP 2014, Berkeley, Calif., 18-21 May 2014*, pp. 212–226.

Ahn, Jaimeen & Alice Oh (2021). Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 533–549. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Benton, Adrian, Raman Arora & Mark Dredze (2016). Learning multiview embeddings of Twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* Berlin, Germany, 7–12 August 2016, pp. 14–19.

Blodgett, Su Lin, Johnny Wei & Brendan O'Connor (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. Melbourne, Australia: Association for Computational Linguistics.

Blue, Ethan, Michael Levine & Dean Nieusma (2014). *Engineering and War: Militarism, Ethics, Institutions, Alternatives*. Morgan and Claypool Publishers.

Bolukbasi, T., K.-W. Chang, J.Y. Zou, V. Saligrama & A.T. Kalai (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in NeurIPS*.

Brennan, Michael, Sadia Afroz & Rachel Greenstadt (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security*, 15(3). Article No. 12.

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

Çarık, Buse & Reyyan Yeniterzi (2022). A Twitter corpus for named entity recognition in Turkish". In *Proceedings of the 13th International Conference on Language Resources and Evaluation,* Marseille, France, 20–25 May 2022, pp. 4546–4551.

Church, Kenneth Ward & Valia Kordoni (2021). Emerging trends: Ethics, intimidation, and the Cold War. *Natural Language Engineering*, 27:379–390.

Czarnowska, Paula, Yogarshi Vyas & Kashif Shah (2021). Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Del Tredici, Marco, Diego Marcheggiani, Sabine Schulte im Walde & Raquel Fernández (2019). You shall know a user by the company it keeps: Dynamic representations for social media users in NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing,* Hong Kong, 3–7 November 2019, pp. 4707–4717.

Delobelle, Pieter, Ewoenam Tokpo, Toon Calders & Bettina Berendt (2022). Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1693–1706. Seattle, United States: Association for Computational Linguistics.

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.

Emmery, Chris, Ákos Kádár & Grzegorz Chrupała (2021). Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics,* Online, 19–23 April 2021, pp. 2388–2402.

Flek, Lucie (2020). Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Online, 5–10 July 2020, pp. 7828–7838.

Floridi, Luciano (2005). The ontological interpretation of informational privacy. *Ethics and*

Information Technology, 7:185–200.

Foulds, James R, Rashidul Islam, Kamrun Naher Keya & Shimei Pan (2020). An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1918–1921, IEEE.

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.

Garg, Sahaj, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi & Alex Beutel (2019). Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 219–226.

Garimella, Aparna, Carmen Banea, Dirk Hovy & Rada Mihalcea (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3493–3498. Florence, Italy: Association for Computational Linguistics.

Gonen, Hila & Yoav Goldberg (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Hardt, M., E. Price & N. Srebro (2016). Equality of opportunity in supervised learning. In *Advances in NeurIPS*, pp. 3315–3323.

Huang, Po-Sen, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama & Pushmeet Kohli (2020). Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 65–83. Online: Association for Computational Linguistics.

Islam, Rashidul, Shimei Pan & James R Foulds (2021). Can we obtain fairness for free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 586–596.

Jernite, Yacine, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Gérard Dupont, Jesse Dodge, Kyle Lo, Zeerat Talat, Isaac Johnson, Dragomir Radev, Somaieh Nikpoor, Jörg Frohberg, Aaron Gokaslan, Peter Henderson, Rishi Bommasani & Margaret Mitchell (2022). Data governance in the age of large-scale data-driven language technology. In *FAccT 2022*.

Jiang, Liwei, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenna Liang, Jesse Dodge, Keisuke Sakaguchi, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini & Yejin Choi (2022). *Can machines learn morality? The Delphi experiment*. arXiv: 2110.07574v2.

Jurgens, David, Yulia Tsvetkov & Dan Jurafsky (2017a). Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 51–57.

Jurgens, David, Yulia Tsvetkov & Dan Jurafsky (2017b). Writer profiling without the writer's text. In *Proceedings of Social Informatics: 9th International Conference, Oxford, UK, 13-15 September 2017*, pp. 537–558.

Kusner, M.J., J. Loftus, C. Russell & R. Silva (2017). Counterfactual fairness. In *NeurIPS*, pp. 4069–4079.

Lee, Kahyun, Mehmet Kayaalp, Sam Henry & Özlem Uzuner (2022). A context-enhanced de-identification system. *ACM Transactions in Computational Healthcare*, 3(1).

Liang, Paul Pu, Chiyu Wu, Louis-Philippe Morency & Ruslan Salakhutdinov (2021). Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pp. 6565–6576, PMLR.

Mahmood, Asad, Zubair Shafiq & Padmini Srinivasan (2020). A girl has a name: Detecting authorship obfuscation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Online, 5–10 July 2020, pp. 2235–2245.

Mahowald, Kyle, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum & Evelina Fedorenko (2023). Dissociating language and thought in large language models: a cognitive perspective. *https://arxiv.org/abs/2301.06627*.

Mubarak, Hamdy, Shammur Absar Chowdhury & Firoj Alam (2022). ArabGend: Gender analysis and inference on Arabic Twitter. In *Proceedings of the 8th Workshop on Noisy User-*

generated Text , Gyeongju, Korea, 16 October 2022, pp. 124–135.

Nadeem, Moin, Anna Bethke & Siva Reddy (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371. Online: Association for Computational Linguistics.

Nguyen, Thong, Andrew Yates, Ayah Zirikly, Bart Desmet & Arman Cohan (2022). Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Dublin, Ireland, 22-27 May 2022, pp. 8446–8459.

Noble, S.U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike & Ryan Lowe (2022). Training language models to follow instructions with human feedback. *https://arxiv.org/abs/2203.02155*.

Qian, Chen, Fuli Feng, Lijie Wen, Chunping Ma & Pengjun Xie (2021). Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5434–5445. Online: Association for Computational Linguistics.

Reddy, Sravana & Kevin Knight (2016). Obfuscating gender in social media writing. In *Proceedings of the 1st Workshop on Natural Language Processing and Computational Social Science,* Austin, Texas, 5 November 2016, pp. 17–26.

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. New Orleans, Louisiana: Association for Computational Linguistics.

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.

Sap, Maarten, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith & Yejin Choi (2020). Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5477–5490. Online: Association for Computational Linguistics.

Schick, Timo, Sahana Udupa & Hinrich Schütze (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Shing, Han-Chin, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III & Philip Resnik (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp. 25–36.

Talat, Zeerak, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell & Adina Williams (2022). On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Seattle, Wash., and Online, 10–15 July 2022, pp. 769–779.

Waseem, Zeerak (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the 1st Workshop on Natural Language Processing and Computational Social Science,* Austin, Texas, 5 November 2016, pp. 138–142.

Waseem, Zeerak, Thomas Davidson, Dana Warmsley & Ingmar Weber (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First*

Workshop on Abusive Language Online, Vancouver, B.C., Canada, 4 August 2017, pp. 78–84.

Waseem, Zeerak & Dirk Hovy (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop,* San Diego, Cal., 12-17 June 2016, pp. 88–93.

Zhang, Brian Hu, Blake Lemoine & Margaret Mitchell (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340.

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, Volume 2 (Short Papers)*, pp. 15–20. New Orleans, Louisiana: Association for Computational Linguistics.

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. Brussels, Belgium: Association for Computational Linguistics.

Zhou, Xuhui, Maarten Sap, Swabha Swayamdipta, Yejin Choi & Noah Smith (2021). Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3143–3155. Online: Association for Computational Linguistics.