Integrating Ethics into the NLP Curriculum

Emily M. Bender
University of Washington
ebender@uw.edu

Dirk Hovy
Bocconi University
dirk.hovy@unibocconi.it

Xanda Schofield
Harvey Mudd College
xanda@cs.hmc.edu

1 Description

As NLP technology becomes more ubiquitous, it has ever more impact on the lives of people all around the world. As a field, we have become increasingly aware that we have a responsibility to evaluate the effects of our research and mitigate harmful outcomes. This is true for both researchers and developers in universities, government labs, and industry. However, without experience of how to productively engage with the many ethical conundrums in NLP, it is easy to become overwhelmed and remain inactive. To raise awareness among future NLP practitioners and prevent inertia in the field, we need to place ethics in the curriculum for all NLP students — not as an elective, but as a core part of their education. Though ethical considerations are achieving new currency in NLP, similar issues have been under consideration for decades, if not centuries, in other fields, and there are robust existing practices for approaching these problems. The difference is that there is no agreed-upon way to engage with them in our field.

Our goal in this tutorial is to empower NLP researchers and practitioners with tools and resources to teach others about how to ethically apply NLP techniques. Our tutorial will present both high-level strategies for developing an ethics-oriented curriculum, based on experience and best practices, as well as specific sample exercises that can be brought to a classroom.[1] We plan to make this a highly interactive work session culminating in a shared online resource page that pools lesson plans, assignments, exercise ideas, reading suggestions, and ideas from the attendees. Though the tutorial will focus particularly on examples for college classrooms, we believe the ideas can extend to company-internal workshops or tutorials in a variety of organizations.

We consider three primary topics with our session that frequently underlie ethical issues in NLP research:

1. **Dual Use**: Learning how to anticipate how a developed technology could be repurposed for harmful or negative results, and designing systems so that they do not inadvertently cause harm.

2. **Bias**: Understanding the different ways in which bias interacts with language data, including over- and under-sampling of different populations as well as the effects of human bias expressed in language; building less biased datasets and debiasing trained models; strategies for matching appropriate training data to a given use case.

3. **Privacy**: Protecting the privacy of speakers/writers of text used in the construction or evaluation of a new NLP technology.

In this setting, a key lesson is that there is no single approach to ethical NLP: each project requires thoughtful consideration about what steps can be taken to best support people affected by that project. However, we can learn (and teach) what kinds of issues to be aware of and what kinds of strategies are available for mitigating harm. To teach this process, we apply and promote interactive exercises that provide an opportunity to ideate, discuss, and reflect. We plan to facilitate this in a way that encourages positive discussion, emphasizing the creation of ideas for the future instead of negative opinions of previous work.

2 Type of tutorial

Introductory. Though this is a topic of importance to the NLP community internally, it relies

[1] The specific exercises we propose include ones that have been field-tested.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6–9
July 5, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

on existing expertise from both pedagogical and philosophical work, and it is not meant to depend on any particular research area of NLP. However, we do believe the content of this workshop also explores questions not fully answered in our field about concrete best practices in the specific context of NLP courses.

A note on interactivity: The proposed format of this tutorial is different from many past introductory tutorials, in the sense that it relies heavily on participation as part of the instruction. However, we believe this is a necessary part of the format of this tutorial for several reasons:

- Because our tutorial is focused on pedagogy, it makes sense to use effective and equitable pedagogical classroom techniques in it. Interactivity through active or cooperative learning (Slavin, 1980; Johnson and Johnson, 2008) and guided discovery-based learning (Alfieri et al., 2011) are proven to enable students to learn more effectively across diverse classrooms, and our design models this.

- The outcome of this tutorial is one focused on training and professional development, which comes with practice. In the same way one might encourage developing a sample neural network in a tutorial on deep learning, we encourage performing steps of educational practice to develop skills to then use in our lives as instructors.

- While there exists literature in ethics pedagogy and ethics in NLP, there do not exist large pools of resources and papers to refer when designing a course, but instead only a small collection of syllabi for ethics in machine learning/NLP courses. An interactive tutorial format allows us to use the learning experiences of our participants as a starting point to construct a more centralized pool of resources from which faculty and educators in NLP can draw.

3 Outline

1. Introduction, background, motivation [10m]

2. Core concepts and terminology, and warm up exercises. [50m] We will have the participants discuss what motivates them and core concepts of ethics and pedagogy that might be useful in the subsequent ideation.

3. Big class exercise I [55m] (5 minutes intro, 35 minutes doing the exercise with the group, 10 minutes talking about how to teach it). The exercises in this set are centered around thinking through how systems behave in the world. There will be a separate exercise for each of the three groups: dual use, bias, and privacy.

   Dual Use A student approaches you because they want to explore gendered language in the LGBTQ community. They are very engaged in the community themselves and have access to data. Their plan is to write a text classification tool that distinguishes LGBTQ from heterosexual language. What do you tell the student?

   Bias Pick an application of speech/language technology, determine what kind of training data is typically used for it (whose language? recorded when/where/how?). Next, imagine real world use cases for this technology. What speaker groups would come in contact with the system? If their language differs from substantially from the training data, what would the failure mode of the system be and what would the real-world impacts of that failure be? How could systems, their training data or documentation be designed to be robust to this kind of problem?

   Privacy Consider a simple Naive Bayes classifier trained on a subset of 20 Newsgroups using word frequencies as features. For five sample messages, could you tell whether or not they were included in the subset? How would you check? How certain could you be?

4. Big class exercise II [55m] (5 minutes intro, 25 minutes refining the exercise, 25 minutes talking about how to teach it). The exercises in this set involve building a system and observing its behavior.

   Dual Use (1) An ACL submission claims to be able to undo ciphers used by dissenters on social media. Who benefits from this? Is it better to release it in a peer-reviewed venue than to not know it? (2) You develop a tool that can detect depression with high accuracy.

Why, or why not, should you release it as an app?

Bias Taking inspiration from Speer (2017b), build a sentiment analysis system over restaurant reviews using different sources of training data for word embeddings. What kind of biases can be observed in system behavior for different types of cuisine? What patterns in language use in the underlying training data are responsible? What kinds of analogous problems can arise in other systems that use word embeddings as input?

Privacy Design a small search engine around an inverted index that uses random integer noise from a two-sided geometric distribution (Ghosh et al., 2012) to shape which queries are retrieved. Analyze how much this changes the search results with different noise levels. Are there systematic changes?

5. Wrap up [20m]: big points, reflections from people, where to find resources and keep talking

4 Prerequisites

This tutorial is meant to be accessible to anyone actively working with NLP and either currently teaching, interested in teaching, or interested in informal instruction outside of university contexts.

5 Reading List

We recommend the following short readings to get a sense of the kinds of issues we will be approaching:

- Dual Use: Ehni 2008
- Bias: Speer 2017a
- Privacy: Coavoux et al. 2018

In addition, we recommend the following papers for a sense of what can be learned from other fields:

- Value scenarios, a technique from value sensitive design: Nathan et al. 2007
- A history of notions of fairness in education and hiring: Hutchinson and Mitchell 2019
- Disparate impact: Feldman et al. 2015

Participants are encouraged to have read at least some of these papers ahead of time, but familiarity with all of them will not be assumed.

6 Instructors

Emily M. Bender
University of Washington
ebender@uw.edu
faculty.washington.edu/ebender

Emily M. Bender is a Professor of Linguistics and Adjunct Professor of Computer Science and Engineering at the University of Washington. Her research interests include computational semantics, grammar engineering, computational linguistic typology, and ethics in NLP. She is the Faculty Director of UW's Professional Masters in Computational Linguistics (CLMS) and has been engaged with integrating ethics into the CLMS curriculum since 2016. She co-organized the first EthNLP workshop. Her first publication in this area is the TACL paper "Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science" (Bender and Friedman, 2018) and she has been an invited speaker at workshops and panels related to ethics and NLP (or AI more broadly) at the Taskar Memorial Event (UW, March 2018), The Future of Artificial Intelligence: Language, Ethics, Technology (Cambridge, March 2019), West Coast NLP (Facebook, September 2019), Machine Learning Competitions for All (NeurIPS, December 2019) and AAAS (Seattle, February 2020).

Xanda Schofield
Harvey Mudd College
xanda@cs.hmc.edu
www.cs.hmc.edu/~xanda

Xanda Schofield is an Assistant Professor of Computer Science at Harvey Mudd College. Her work focuses on the practical aspects of using distributional semantic models for analysis of real-world datasets, with problems ranging from understanding the consequences of data pre-processing on model inference (Schofield and Mimno, 2016; Schofield et al., 2017) to enforcing text privacy for these models (Schein et al., 2018). She also is interested in pedagogy at this intersection, having co-developed a Text Mining for History and Literature course at Cornell University with David Mimno. She is currently focusing pedagogical efforts on how to introduce considerations of ethics and bias into other courses such as Algorithms.

Dirk Hovy
Bocconi University
dirk.hovy@unibocconi.it
www.dirkhovy.com

Dirk Hovy is an Associate Professor of Computer Science in the Department of Marketing at Bocconi University in Milan, Italy. His research focuses on how social dimensions influence language and in turn NLP models, as well as on questions of bias and fairness. He strives to integrate sociolinguistic knowledge into NLP models to counteract demographic bias. Dirk has written on ethics and bias in NLP (Hovy and Spruit, 2016), co-organized two editions of the EthNLP workshops and one of the abusive language workshop, and was an invited speaker on panels on ethics at NAACL 2018 and SLT 2018. He is teaching a related tutorial (on ethics and biases) at CLiC-IT in November 2019.

References

Louis Alfieri, Patricia J Brooks, Naomi J Aldrich, and Harriet R Tenenbaum. 2011. Does discovery-based instruction enhance learning? *Journal of educational psychology*, 103(1):1.

Emily M. Bender and Batya Friedman. 2018. Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6.

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. 2018. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10. Association for Computational Linguistics.

Hans-Jörg Ehni. 2008. Dual use and the ethical responsibility of scientists. *Archivum immunologiae et therapiae experimentalis*, 56(3):147.

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

Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. 2012. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598. Association for Computational Linguistics.

Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of FAT* 19: Conference on Fairness, Accountability, and Transparency (FAT* 19)*, volume abs/1811.10104, New York. ACM.

Roger T Johnson and David W Johnson. 2008. Active learning: Cooperation in the classroom. *The annual report of educational psychology in Japan*, 47:29–30.

Lisa P. Nathan, Predrag V. Klasnja, and Batya Friedman. 2007. Value scenarios: A technique for envisioning systemic effects of new technologies. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*, pages 2585–2590. ACM.

Aaron Schein, Zhiwei Steven Wu, Alexandra Schofield, Mingyuan Zhou, and Hanna Wallach. 2018. Locally private bayesian inference for count models. *arXiv preprint arXiv:1803.08471*.

Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Robert E Slavin. 1980. Cooperative learning. *Review of educational research*, 50(2):315–342.

Robyn Speer. 2017a. Conceptnet numberbatch 17.04: better, less-stereotyped word vectors. Blog post, https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/, accessed 15 January 2019.

Robyn Speer. 2017b. How to make a racist AI without really trying. Blog post, http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/, accessed 15 January 2019.