A Crash Course on Ethics for Natural Language Processing

Annemarie Friedrich[1] **Torsten Zesch**[2]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]Language Technology Lab, University Duisburg-Essen, Germany
annemarie.friedrich@de.bosch.com
torsten.zesch@uni-due.de

Abstract

It is generally agreed upon in the natural language processing (NLP) community that ethics should be integrated into any curriculum. Being aware of and understanding the relevant core concepts is a prerequisite for following and participating in the discourse on ethical NLP. We here present ready-made teaching material in the form of slides and practical exercises on ethical issues in NLP, which is primarily intended to be integrated into introductory NLP or computational linguistics courses. By making this material freely available, we aim at lowering the threshold to adding ethics to the curriculum. We hope that increased awareness will enable students to identify potentially unethical behavior.

1 Motivation and Overview

The recent sharp rise in the capabilities of natural language processing (NLP) methods has led to a wide application of language technology, influencing many aspects of our daily lives. As a result, NLP technology can have considerable real-world consequences (Vitak et al., 2016; Hovy and Spruit, 2016). While language technology has the aim of supporting humans, mis-use of data or abuse of subjects, mis-representation, or direct harm are some of numerous potentially critical problems. Hence, it is crucial for NLP researchers, developers, and deciders to be aware of the social implications of deploying a particular piece of language technology. They should be able to analyze the degree to which a setup conforms with *ethical* principles, which guide what is considered 'right' or 'wrong' (Deigh, 2010), and be aware of the potentially harmful side even of components developed with the aim of supporting humans.

As a consequence, it is crucial to start early and embed ethics into the NLP curriculum to be taught along with the technological and linguistic material in an interactive manner (Bender et al., 2020). While many teachers are very open to the topic area as demonstrated by the lively participation in events such as the 2020 Workshop on Integrating Ethics into the NLP Curriculum (Bender et al., 2020), ethics for NLP also constitutes a very broad topic area where even teachers might not feel knowledgeable enough. We argue that many more lecturers would include ethics if there existed some freely available **ready-made** teaching material that can be easily integrated in existing courses, or that could serve as a starting point for designing a more in-depth course. In this paper, we describe such a material, which is primarily intended to be integrated into introductory NLP courses.

Our "crash course" does not claim exhaustiveness and aims at breadth rather than depth, intending to give a good overview of the field and highlighting potential issues. The main focus is to enable students to behave ethically in their future research and work careers, and to continue their own research and reading. To be able to participate in the discourse on ethical issues in NLP, the first step is to learn about the important **concepts** and **terminology**. The material also provides numerous suggestions for in-class **discussions** and exercises with the aim of gaining a deeper understanding.

The material consists of **extensively commented slides and suggestions for practical exercises**. The comments and lists of references serve both for teacher preparation and as a script for students. The teaching material is freely available under CC-BY-SA from our website.[1] Finally, both ethics and NLP are dynamic fields which may require updating views, beliefs, opinions, and theories. We therefore welcome continuous feedback regarding and contributions to the teaching material.

Benefit of this Course. The main difference versus existing freely available material is that our crash course is **short**, **self-explanatory** in the con-

[1] https://gscl.org/en/resources/ethics-crash-course

Proceedings of the Fifth Workshop on Teaching NLP, pages 49–51

text of the provided comments such that other lecturers can easily work with it, and contains many **linked exercises**. The Markkula Center for Applied Ethics at Santa Clara University offers a slide set with a crash course on ethics along with materials for discussions focusing of ethics in the general area of technology.[2] Our intention is very similar, but the focus is on NLP-specific issues. We found most existing freely-available courses on ethics in NLP to go into depth and usually span an entire semester.[3] In addition, we are aware of several recent tutorials at ACL venues. Most similar to our materials is the tutorial on socially responsible NLP by Tsvetkov et al. (2018), which is a condensed version of a full-semester class.[4] Our course differs from this tutorial in length and by mostly concentrating on NLP-specific examples. Chang et al. (2019) focus on sub-topics of ethical NLP such as fairness and mitigating bias.[5] Bender et al. (2020) have started a collection of pointers to existing materials, as well as general pointers for teaching ethics for NLP.[6] This collection has been a very valuable starting point for our own research.

Ethical Considerations. Admittedly, a potential issue is that teachers could just "check off" the topic by using our material without deeper engagement, individual reading or reasoning. However, we believe that having a good starting point will actually lead to both teachers and students doing more research on this subject, and our companion material emphasizes the benefit of digging deeper.

2 Course Description

Format. The crash course consists of an interactive lecture accompanied by practical exercises. The slides are available as extensively commented Google slides, which can be easily adapted and exported into a number of common formats. Exercises consist of reading assignments, examples and case studies that serve as the basis for pair or group discussions, or essays if submitting written material is essential, e.g., for grading purposes.

Learning Goals. After the crash course, students will have acquired a basic understanding of the relevant terminology and concepts. They should understand that there are different ethical theories at interplay with the field of NLP which are currently developing best practices for ethical conduct and systems (Prabhumoye et al., 2019). As a result, they should be able to critically reflect the on-going discourse in the community and, last but not least, have acquired the basis for behaving ethically in their own work. It is not our goal to provide 'ultimate' definitions of the concepts and we strongly advise lecturers not to pose exam questions aiming at memorizing definitions. Instead, the learning goals could be tested in the form of group presentations or written essays.

The predominant principle behind the design of our crash course is **activation.** Besides giving clear descriptions of relevant concepts and terminology, we believe that a sensitivity for ethical issues can only be achieved by actively thinking about problems. We hence provide discussion questions for most topics covered in the crash course. We strongly recommend taking the time for short pair- or small-group discussions on these questions before further discussion in the plenum.

Topics Covered. Our course aims at providing a good overview of the field, offering references as starting points for deeper research. We consider our teaching materials to be a 'living document' that will be updated or extended continuously. Topics covered in our first version include, among others, bias, fairness, privacy, and analyzing NLP use cases or methods from different ethical perspectives. For an up-to-date overview of the course content, please refer directly to the material.

3 Discussion

Our stated goal is to inform a broader public about the on-going discourse about ethics in NLP, and educate future NLP researchers, developers and deciders about an ethical approach to NLP research technology. We hope that our materials will be of benefit not only in university classrooms, but also in other settings such as reading groups or industrial meet-ups. We hence publish our resources under the CC-BY-SA 4.0 license,[7] which, under the conditions of stating the source and redistribution under the same license, allows copying, redistributing, adapting and mixing the material in any medium or format.

[2] https://www.scu.edu/ethics-in-technology-practice

[3] See, e.g, http://demo.clab.cs.cmu.edu/ethical_nlp2020/#readings, http://faculty.washington.edu/ebender/2017_575

[4] Materials are also publicly available at https://sites.google.com/view/srnlp/.

[5] http://web.cs.ucla.edu/~kwchang/talks/emnlp19-fairnlp

[6] https://aclweb.org/aclwiki/Notes_on_Teaching_Ethics_in_NLP, https://aclweb.org/aclwiki/Ethics_in_NLP

[7] https://creativecommons.org/licenses/by-sa/4.0/

Acknowledgments

We thank Dirk Hovy for sharing his materials on ethics and for his feedback on this crash course. We also thank Ronja Laarman-Quante, Sophie Henning, Heike Adel, Andrea Horbach, and the anonymous reviewers for their valuable feedback.

References

Emily M. Bender, Dirk Hovy, and Alexandra Schofield. 2020. Integrating ethics into the NLP curriculum. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9, Online. Association for Computational Linguistics.

Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

John Deigh. 2010. *An Introduction to Ethics*. Cambridge University Press.

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, Berlin, Germany. Association for Computational Linguistics.

Shrimai Prabhumoye, Elijah Mayfield, and Alan W Black. 2019. Principled Frameworks for Evaluating Ethics in NLP Systems. *arXiv preprint arXiv:1906.06425*.

Yulia Tsvetkov, Vinodkumar Prabhakaran, and Rob Voigt. 2018. Socially responsible NLP. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 24–26, New Orleans, Louisiana. Association for Computational Linguistics.

Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 941–953, New York, NY, USA. Association for Computing Machinery.