Socially Responsible Natural Language Processing

Yulia Tsvetkov
Carnegie Mellon University
ytsvetko@cs.cmu.edu

Vinodkumar Prabhakaran
Google
vinodkpg@google.com

Rob Voigt
Stanford University
robvoigt@stanford.edu

ABSTRACT

As language technologies have become increasingly prevalent in analyzing online data, there is a growing awareness that decisions we make about our data, methods, and tools often have immense impact on people and societies. This tutorial will provide an overview of real-world applications of Natural Language Processing technologies and their potential ethical implications. We intend to provide the researchers with an overview of tools to ensure that the data, algorithms, and models that they build are socially responsible. These tools will include a checklist of common pitfalls that one should avoid, as well as methods to mitigate these issues. Issues of bias, ethics, and impact are often not clear-cut; this tutorial will also discuss the complexities inherent in this area.

CCS CONCEPTS

• **Social and professional topics** → **Codes of ethics**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

socially responsible computing; ethical considerations in computing; ethics in natural language processing; bias in data and models; textual data; privacy; human-generated data

ACM Reference Format:
Yulia Tsvetkov, Vinodkumar Prabhakaran, and Rob Voigt. 2019. Socially Responsible Natural Language Processing. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3308558.3320097

1 TOPIC AND RELEVANCE

In this tutorial we will discuss the philosophical foundations of ethical research along with state-of-the art machine learning fairness methods, with a particular focus on language data, algorithms, and applications. Major discussion topics include:

- Foundations of ethics in NLP: what is ethics, history, medical and psychological experiments, ethical decision making, a case study that identifies ethical pitfalls in working with computational models that can affect human lives.
- Working with people and human biases in language data: IRB and human subjects, working with crowdsourced workers, biases inherent in language data, privacy considerations.
- Methods to detect and mitigate biases in NLP models.

With a half-day tutorial and many possible topics, we do not plan an exhaustive treatment of this material. One central goal is to raise awareness for participants of the relevant issues, so that when they return to their research they will be more able to notice ways that ethical issues may play out in different contexts. To achieve this goal, we will aim for a "T-shape" in terms of breadth and depth: to briefly mention a number of core ethical questions, and then to drill down into a few particular case studies to see how these issues play out in real research settings.

As computational linguists, we often see the internet as a vast, interconnected set of "documents," which are largely composed of text written in natural language. Many WWW participants do work that touches upon working with language on the internet directly or indirectly, but even for those who do not work with language, issues of social bias are likely to find substantial relevance. One issue we hope to discuss with the WWW community in detail is the variety of sampling biases in internet research: often data accessibility, the availability of APIs, and the terms of service for web platforms influence what research is possible to do, and we would encourage a discussion of strategies to acknowledge and mitigate these issues.

2 AUDIENCE

We expect participants from a wide array of backgrounds, including researchers, engineers, and end users of WWW technologies. No prior experience with NLP/ML is required, but we believe that our tutorial will most benefit those who are currently using NLP or are intending to use NLP in the near future in their research/products.

3 PREVIOUS EDITIONS

The first iteration of this tutorial (in the half-day format) was given at NAACL 2018: https://sites.google.com/corp/view/srnlp/tutorial-naacl18. Link to the slides are at the bottom of the page. We have updated the content to include more recent work in this area.

4 RELEVANT LINKS AND COURSES

We maintain a web page relevant to the tutorial at https://sites.google.com/view/srnlp.

Additional relevant courses in the intersection of Ethics and NLP:

- Emily Bender at Univ. of Washington: http://faculty.washington.edu/ebender/2017_575/
- Graham Hirst at Univ. of Toronto: http://www.cs.utoronto.ca/~gh/cscD03/index.shtml
- Yulia Tsvetkov and Alan W Black's at CMU: http://demo.clab.cs.cmu.edu/ethical_nlp/

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308558.3320097