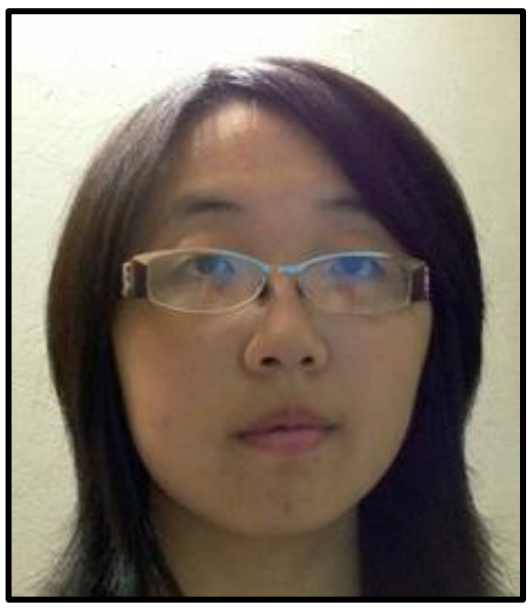
Language Technologies Institute

Thesis Defense

Monday, August 9, 2021| 3:00 pm



Evaluating and Recontextualizing the Social Impacts of Moderating Online Discussions

Qinlan Shen

Abstract

In response to recent surges in abusive content in online spaces, large social media companies, such as Facebook, Twitter, and Reddit, have been pressured both legally and socially to strengthen their stances against offensive content on their platforms. While the standard practice for addressing abusive content is for human moderators to review whether content is appropriate, the vast scale of online content and psychological toll of abusive material on moderators has led to growing interest in natural language processing in developing technologies to aid in the moderation of offensive language. However, while there has been steady progress on the development of models centered on classifying offensive texts, there is limited consensus over what abusive language is and how NLP models can address practical issues within online moderation. In the complex sociotechnical systems where content moderation takes place, the answers to the questions of "what is abusive language?" and "how should language technologies be used to address abusive language?" can have a major impact on the participation experiences of users in online platforms. Research in online moderation from other disciplines, such as human-computer interaction, platform design, and law, often addresses these social consequences by taking a more interaction-focused view of the problem of moderating abusive language. However, when evaluating moderation issues at scale, these studies of interaction often end up relying on simplified approaches for considering sociolinguistic issues in online communities.

In this thesis, my goal is to bridge the gap between the language-focused view of content moderation from NLP and the interaction-based view from platform design in two directions. In the first direction, I contribute to the more interactional view of moderation by introducing and applying more sophisticated conceptions of language to aid in the evaluation of social impacts of moderation strategies using quantitative methods at scale. Under this evaluation paradigm, I demonstrate the use of NLP techniques in measuring the social impacts of moderation strategies through three case studies over different online communities at different levels of impact. In the other direction, I contribute to the more language-focused view of offensive language moderation by leveraging insights from social theories and the study of online communities to better contextualize what linguistic variation looks like in online spaces. Under this contextualization paradigm, I introduce an examination of how to operationalize descriptive linguistic norm differences across political subcommunities on Reddit. Based on these analyses, I discuss broader implications of my work for platform design and language technologies and reflect on future directions in both disciplines that may contribute to addressing abusive language in online spaces.

https://drive.google.com/file/d/1f2YdW8T7ByIUKK6IEc5Ep4XRSdpm5rpS/view

COMMITTEE:

Carolyn P. Rosé
(Chair)

Yulia Tsvetkov

Geoff Kaufman

David Jurgens,
(University of Michigan)

Cliff Lampe,
(University of Michigan)