Language Technologies Institute

Thesis Proposal

GHC 4405| Monday, February 24th | 11:00 am

Evaluating and Recontextualizing the Social Impacts of Offensive Language Moderation

Qinlan Shen

COMMITTEE:

Carolyn Rosé, (chair)

Yulia Tsvetkov

Geoff Kaufman

David Jurgens,
(Univ. of Michigan)

Cliff Lampe,
(Univ. of Michigan)

Abstract

Recent surges in abusive content on social media have led to increased interest in NLP research in developing technologies to aid in moderating offensive language online. While there has been steady progress in developing models for detecting offensive language, there is little consensus over what problems these technologies should address and limited understanding of their social consequences. Researchers in other fields, such as law, platform design, and human-computer interaction have taken a more interaction-based approach in considering issues in offensive language moderation by examining questions like the social impacts of interventions. However, large-scale studies of moderation impacts often take a simplified view of language-related issues.

This thesis attempts to address the gap between the text-centered view of the offensive language problem in NLP and the interaction-centered view in platform design in two ways. First, we develop and apply more nuanced techniques for operationalizing offensive language to allow for more consistent evaluation of the impacts of moderation strategies at scale. Under our evaluation-based paradigm, we present case studies of moderation strategies on Ravelry and Reddit to highlight difficult social issues to consider when thinking about moderation strategies. In our second paradigm, we propose to use interaction to reconsider how we define offensive language in online communities through contextualization. Our goal is to examine the role that interactional context plays in making an offensive language judgment. We will consider the tasks of identifying escalation, norm violation, and reaction across several different communities/platforms to investigate to what extent "offense" is built into a text vs. the interactions around it.

https://drive.google.com/file/d/16WLiy5TO0uY6iKtrGQM1GCFrtoaZM0J/view