Predicting and Addressing Hostile Social Media Content Algorithmically

Faculty supervisors:
  - Libby Hemphill
  - David Jurgens

Anti-social, destructive behaviors online (e.g., online harassment, cyberbullying, hostility, trolling) are increasingly problematic. This project focuses on a particular type of destructive content - hostile comments - because they feature prominently in problems such as harassment, incivility, and cyberbullying. The project aims to predict whether particular comments lead to escalations of hostility and to evaluate automated approaches for preventing escalations or triggering de-escalations.

This project develops machine learning approaches to forecast imminent hostility threats in online social networks. Such forecasts can then be used by individuals and communities to take timely actions to avoid escalation and reduce further harm. Additionally, using these forecasts, this project investigates automated de-escalation techniques to help communities proactively respond to potential threats.

The specific objectives are as follows:
1. Understand where and how hostility currently occurs.
2. Develop machine learning algorithms to detect flashpoints, the moments at which online interactions turn harassing or dangerous.
3. Implement a cross-platform application that predicts imminent hostility threats in real-time by monitoring a user’s online social interactions.
4. Develop automated de-escalation techniques best-suited to social context.

We have an existing computational model that reliably detects hostility in Instagram comments, and work in the summer of 2018 will include one or more of the following:

- Discover cultural norms around hostility in a particular space
- Measure when labels (e.g., hate speech) apply and/or drift among contexts
- Automatically learn phrases that de-escalate hostilities
- Uncover the social structure of escalation
- Measure efficacy of de-escalation strategies in real-world randomized trials

Interested students should have strong Python skills and at least some experience working with social media data and/or building machine learning models.
Faculty supervisors will meet weekly with the student. We expect students to work well independently but will happily help with all aspects of the project including troubleshooting code and understanding model output.