ILLINOIS INSTITUTE OF TECHNOLOGY
impact

Nayar Prize II Phase I Quarterly Progress Report (Q3)

July 2017

Project: **Cyberbullying Early Warning and Response System**
Team: **Aron Culotta** and **Libby Hemphill**

Overview

The goal of the Cyberbullying Early Warning and Response System project is to develop software tools to forecast imminent cyberbullying threats and vulnerabilities in online social networks that individuals and communities can use to avoid escalation. This quarter we focused on improving the prediction models by using comment-level data. We also interviewed parents of teenagers to check our assumptions about how cyberbullying unfolds and to elicit feature requests for our warning/intervention system.

Progress Summary

Model Specification: Computer science doctoral student Ping Liu leads our model building efforts. During Q3 he added features from databases of hate speech and profanity to the model and uses all prior comments to predict whether a future comment will be hostile. He also identified cases where the prediction model and Turkers disagree about whether a comment indicates cyberbullying. He also created a set of features based on synonyms and a chatbot that generates comments in reply to Instagram comments.

Data Annotation: Josh Guberman leads our data annotation efforts and iterated on training and annotation HITs through Amazon Mechanical Turk.

As of June 30, 2017, we had 13,266 Instagram comments labeled in which 2727 exhibited features of cyberbullying. Of those 2727, only 70 comments contained physical threats. The relative rarity of these comments makes it hard to get enough data to effectively train classifiers to detect them, and we are working on ways to oversample threats. We also analyzed the cases in which our model mislabeled training data and found that conversations about celebrities and the use of the “tears of joy” emoji are especially confusing to our labelers and the model.

Stakeholder Interviews: Libby interviewed six parents of 12-18 year olds about their definitions of cyberbullying and how they would like to respond to incidents. Her respondents were nearly uniform in their definitions—two or more kids who know each other offline escalating social conflict through social media that often culminates in a physical altercation—and in their desires for a passive monitoring system that would alert them to potential incidents. Parents were not interested in frequent notifications or updates but rather requested a text message or email notification if something concerning was happening on their child’s account. They wanted a prompt to talk to their kids but not necessarily details about the potential incident. These definitions and responses were in line with what we heard from law officers and school officials in Q1 and will inform the user interface design in Q4.

Plans for Q4

In Q4, we continue the work from Q3 by improving the model using additional features such as the shape of the distribution of hostile comments, the timing of comments, and additional labeled data. We are also designing the user interface for the user-facing components of the warning system. Next steps: The team will iterate to refine accuracy of model and forecasting ability and then implement the user interface reflecting input from parent interviews.