Data Justice Lab Seminar Series:
Detecting Online Hate Speech Using Both Supervised and Weakly Supervised Approaches

Time: 10-11am, September 29, 2021 (US central time)
Zoom Meeting ID: 946 2184 9142. Passcode:133541
Direct Zoom link:
https://tamu.zoom.us/j/94621849142?pwd=amxBZFk0NWdoYldSNkw5UCtUcnFxdz09
Faculty host: Lu Tang, Dept of Communication

Dr. Ruihong Huang
Associate Professor
Department of Computer Science & Engineering
Texas A&M University (TAMU), College Station.

Dr. Huang received her PhD in computer science at the University of Utah. She joined TAMU in Fall 2015 after she completed a Postdoc at Stanford University. Her research is mainly on Natural Language Processing (NLP), focusing on information extraction, discourse analysis and semantics. Her research interests also include detecting and mitigating the effects of unwanted language and information, such as media bias, online abusive language or deceptive information. She is a recipient of NSF CAREER award.

Abstract
Online user generated text is laden with hateful content. Context accompanying a hate speech text is useful for identifying hate speech, which however has been largely overlooked in existing datasets and hate speech detection models. We created an annotated corpus of hate speech with context information well kept. We further implemented supervised hate speech detection models that incorporate context information, a logistic regression model with context features and a neural network model with learning components for context. Then, to address various limitations of supervised hate speech classification methods including corpus bias and huge cost of annotation, we proposed a weakly supervised two-path bootstrapping approach for online hate speech detection that leverages large-scale unlabeled data and requires minimal human supervision. Applying this model on a large quantity of tweets collected before, after, and on 2016 election day reveals motivations and patterns of inflammatory language.

TEXAS A&M
Institute of
Data Science