Machine Learning for Detecting Detrimental Online Social Behavior

Bert Huang (with Elaheh Raisi)
Department of Computer Science, Virginia Tech

The Power and Perils of the Social Web

The modern Internet amplifies our ability to communicate.
Theoretically, people can form relationships and communities regardless of location, race, ethnicity, or gender.
But the amplification of social interaction also includes detrimental behaviors such as harassment and cyberbullying.
These behaviors are dangerous. They create harmful, threatening environments and represent a public health hazard.

Cyberbullying

The Cyberbullying Research Center defines cyberbullying as "willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices."
According to stopbullying.gov, victims of bullying are more likely to experience depression and anxiety, health issues, and decreased academic performance.
Cyberbullying is persistent and public. It is not bounded by location or time of day, and it can be instigated anonymously and at large scale.

Social Media Analytics

Social media is a major catalyst in the growth of data science.
Data-driven business is thriving, using human data to improve marketing, content recommendation, user retention, and advertisement allocation.
Computational methods in graph mining and relational learning extract models and analytics by considering the complex social structure of online communication.
Can we retarget some of the machine learning methods developed for business applications toward mitigating the harm from cyberbullying?

Participant-Vocabulary Consistency Model

The model attributes each user with a bully score and a victim score and each word with a harassment word score.
It then minimizes the discrepancy between participant score and vocabulary score for each social media post using alternating least squares.

Experiments

Data: We collected 293k tweets from 180k users, and we used the Ask.fm data collected by Hosseinmardi et al. (2014).
Bullying vocabulary: We created a list of curse words and bigrams from noswearing.com.
Evaluation: We provide the learning algorithm with half of the vocabulary. These ROC curves illustrate its performance.
Comparisons: we compare against direct co-occurrence and dynamic query expansion based on document frequency.

Discussion

We created an initial algorithm for training a model of cyberbullying from weakly supervised data.
The participant-vocabulary consistency model is able to recover known bullying indicators and discover new indicators (not shown here due to explicit content). We evaluated these qualitatively and were impressed.
We are currently working on creating a more formal probabilistic model for bullying to more robustly incorporate noise and uncertainty.

My group is also working on general-purpose methods for scalable learning and inference in structured models.
We also target other applications that have potential societal benefits, such as analysis of large-scale pedagogical data from MOOCs and other learning management systems.
Read about our research at http://berthuang.com.