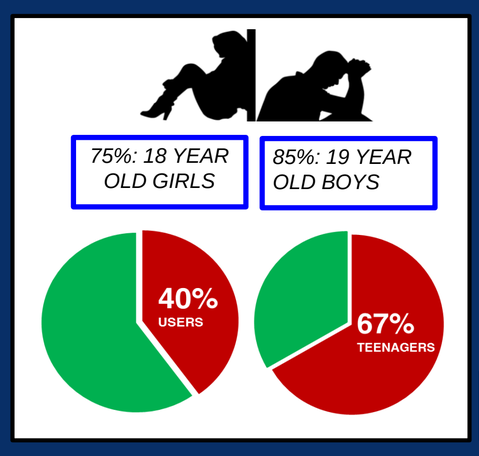
Context-Aware Harassment Detection on Social Media

Amit P. Sheth, Krishnaprasad Thirunarayan, Valerie L. Shalin

Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis), Wright State University, Dayton OH, USA

http://wiki.knoesis.org/index.php/Context-Aware_Harassment_Detection_on_Social_Media

Problem: Harassment on Social Media & Its Impact



Challenges

- *Little past research on automatic detection of harassment on social media.*
- *Existing work on harassment detection predominantly uses machine learning, relying on message content while ignoring message context.*
- *Social networking sites, Facebook, Twitter, and YouTube, have not yet developed effective techniques against harassment.*

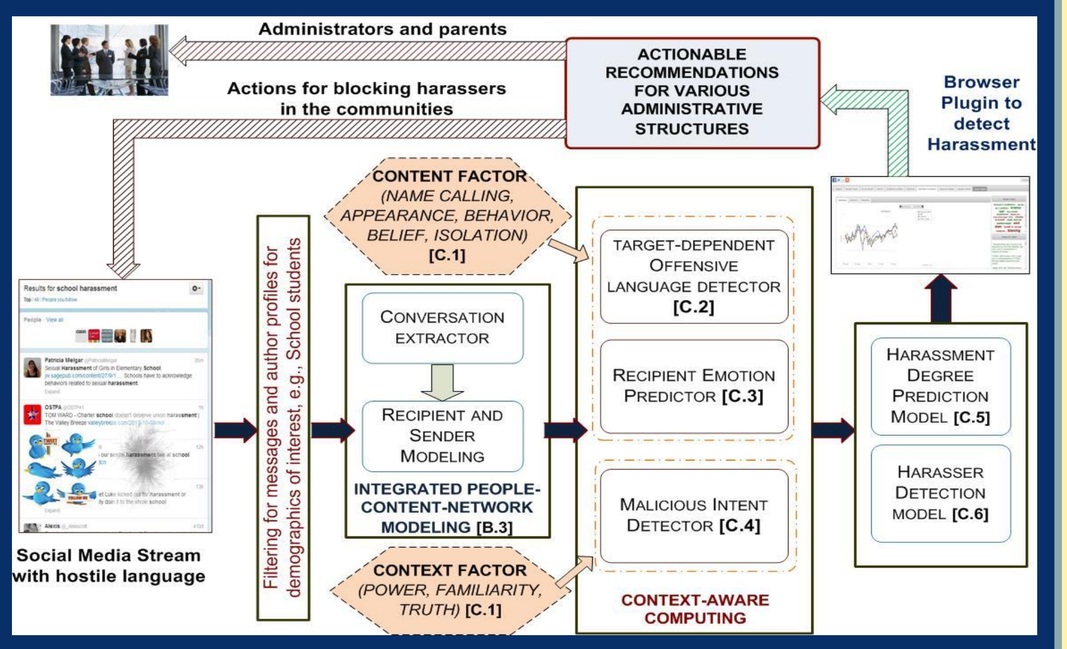
Project Objectives

- **Identify** the language based target-dependent offensiveness/negativity of a message
- **Recognize** sender malice from an intent perspective
- **Predict** message harm from an emotion perspective
- **Detect harassing** social media accounts automatically
- **Evaluate algorithm** quality and generality for both school and workplace settings
- **Provide an alert service** of potential harassment messages for parents to facilitate intervention
- **Educate teenagers** regarding social media harassment, including its characteristics, the associated prohibitions and penalties, and prevention strategies

Some Completed Tasks

- **Filtering author profiles**: To study cyber bullying among school children, we identify tweets from specific high school locations.
- **Identify Harassment Content Factors**: We distinguish between two classes of harassment: (a) threat to physical survival and (b) damage to social capital.
- **Identify Comprehensive Harassment Context Factors**: power, familiarity and truth, explicit and implicit entities.
- **Multimodal transfer**: Extend the analysis of tweets to new sources of data: microblogs, question*answers, blogs and forum conversations.

Proposed Approach Overview



Sample of Ongoing Research: Word Embedding Based Data Augmentation for Harassment Detection: Lexical Substitution Approach

Research Motivation:

- All existing machine learning based approaches to the harassment detection problem rely merely on the labeled instances in the training dataset.
- The short, incomplete and lexically diverse nature of training instances as well as the sparsity of harassing messages suggest the need for *data augmentation.*

Solution:

- Use various word embedding based approaches for augmentation of labeled training instances.
- Use a *novel augmentation* approach using the dependency based word embedding model, to replace each word in the training dataset with more semantically and syntactically meaning-preserving substitute.

Sample of Ongoing Research: Mining YouTube Transcripts for Detecting Harassing Videos

Research Motivation:

- Text messages are not the only form of social media harassment. Harassment also occurs using images and videos.
- This exercise tests our recipient emotion analysis and our target-based notion of offensiveness-- we must distinguish between *offensiveness of the video* and offensiveness of the commenter.

Solution:

- Only “top-level” comments exclusively concern video content (as other comments may be arguments among the commenters).
- Presently we are using n-grams and sentiments of “top-level” comments, and the metadata associated with the YouTube video to determine whether the video is offensive.

Interested in meeting the PIs? Attach post-it note below!

National Science Foundation
WHERE DISCOVERIES BEGIN

The 3rd NSF Secure and Trustworthy Cyberspace Principal Investigator Meeting
January 9-11, 2017
Arlington, Virginia

NSF Award#: CNS 1513721

Ending the Bullying Nightmare
KNOESIS
WRIGHT STATE UNIVERSITY