BullyBlocker

Evaluating a Model to Predict Cyberbullying in Social Networks

Advisor: Yasin N. Silva | Deborah L. Hall   Students: Rusty Conway | Tim Strom | John Spokes | Carmen Sanchez | Victoria Delgadillo | Tom Schenk

ASU New College — ARIZONA STATE UNIVERSITY
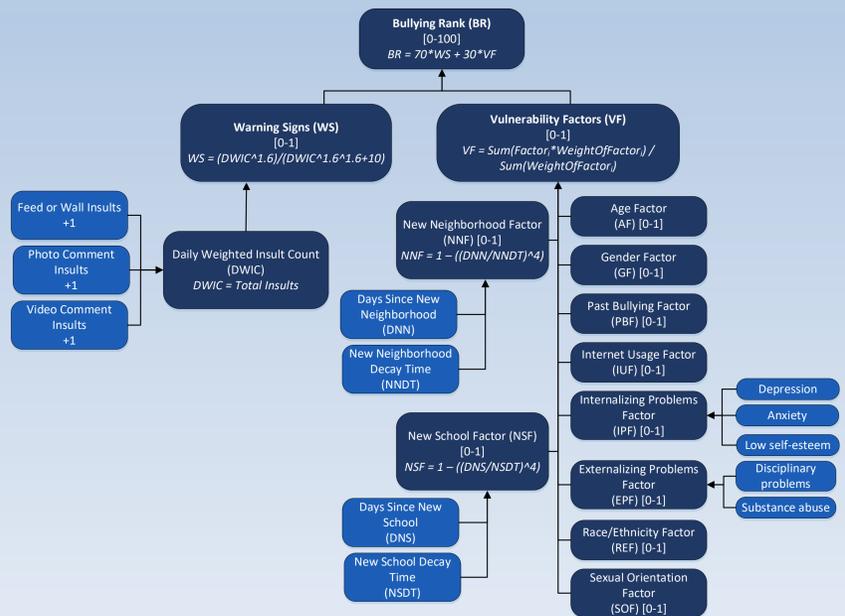
NSF — NCUIRE

Overview

Problem

- Cyberbullying is the use of online digital media to communicate false, embarrassing, or hostile information about another person. It is the most common online risk for adolescents and well over half of young people do not tell their parents when it occurs.

Contribution of Present Research

- Design of a holistic model that: (1) integrates crucial findings from psychological research on predictors of cyberbullying, as well as the relative strength of various predictors and temporal aspects of cyberbullying, and (2) considers streams or bursts of messages in conjunction with information from adolescents' social media profiles to estimate cyberbullying risk

- Introduction of: (1) a module providing a customized list of parent/victim resources using information pertaining to the nature of specific cyberbullying instances, and (2) a novel framework for evaluating cyberbullying identification models using simulated social networks containing content from real world instances of cyberbullying

- Identification of challenges and opportunities to integrate results from psychology and social network data analysis to address a social problem

- Public and no-cost availability of: (1) the BullyBlocker app in the Apple App Store, (2) source code for the evaluation framework, and (3) real-world datasets for use by other researchers to generate simulated social network interactions
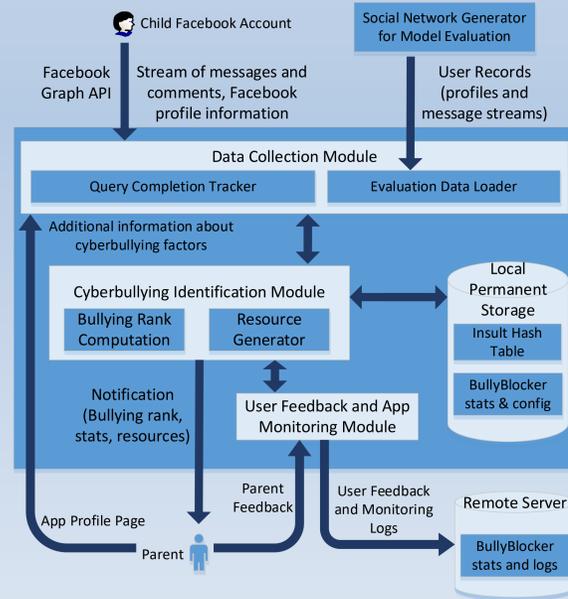
Risk Factors

$$BR = 70*WS + 30*VF$$

Bullying Rank (BR) [0-100]

Warning Signs (WS) [0-1]
$$WS = (DWIC^1.6)/(DWIC^1.6+10)$$

Vulnerability Factors (VF) [0-1]
$$VF = Sum(Factor_i*WeightOfFactor_i) / Sum(WeightOfFactor_i)$$

Daily Weighted Insult Count (DWIC)
$$DWIC = Total Insults$$

New Neighborhood Factor (NNF) [0-1]
$$NNF = 1 - ((DNN/NNDT)^4)$$

New School Factor (NSF) [0-1]
$$NSF = 1 - ((DNS/NSDT)^4)$$

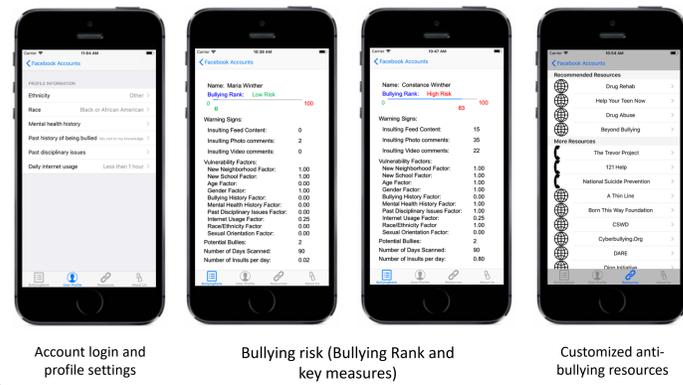| Factor	Details	Weight
New School	# days in a new school	0.10
New Neighborhood	# days in a new neighborhood	0.10
Age	Applied if value is 11-16	0.04
Gender	Applied if value is female	0.12
Race/ Ethnicity	Applied if race is non-white or if ethnicity is Hispanic/Latino	0.02
Sexual Orientation	Applied if self-identified as LGBTQ	0.29
Past Bullying	Applied if user experienced bullying in last 1 month, 1-2 months, more than 2 months	0.42
Daily Internet Use	Considers ranges <1h, 1h-3h, 4h-6h, >6h	0.17
Internalizing Problems	Considers history of depression, anxiety, low self-esteem	0.28
Externalizing Problems	Considers history of disciplinary issues or substance use	0.21

Design

- BullyBlocker analyzes adolescents' interactions with their social network to identify:
  - Warning signs: e.g., number of insulting messages and embarrassing pictures.
  - Vulnerability factors: e.g., race, age, gender, sexual orientation, having recently moved to a new neighborhood or a new school, past bullying history, frequency of internet use, internalizing problems (depression, anxiety, and low self-esteem), and externalizing problems (disciplinary problems and substance abuse).
- Bullying Rank: Estimates the probability of a minor experiencing cyberbullying.
- The computed Bullying Rank is returned to the parent or guardian of the minor.
- The Bullying Rank Is divided into three normalized levels of risk intensity: Low[0-33], Medium[34-66], High[67-100]
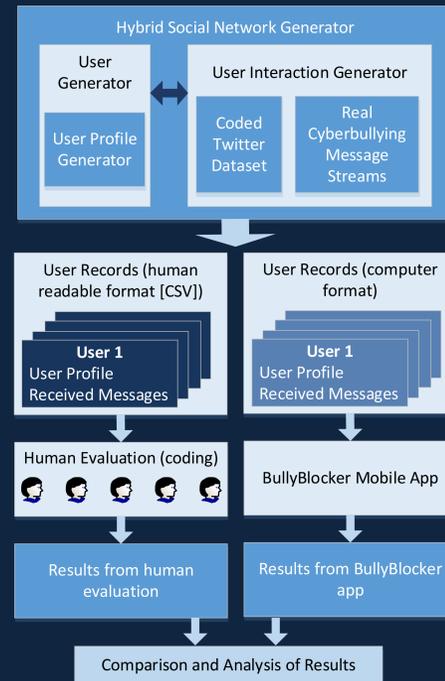
Architecture

Identification and Notification

Account login and profile settings

Bullying risk (Bullying Rank and key measures)

Customized anti-bullying resources

Evaluation Framework

The proposed evaluation framework uses a hybrid social network generator to create realistic datasets provided as input for the app-later used in the human evaluation phase.

Hybrid Social Network Generator. This component generates a test social network composed of synthetic users and real-world interactions. The *User Interaction Generator* module outputs a set of N users, their profile information, and produces a set of interactions among the users. The goal of the module is to create message sequences like those found in real-world social networks. The generator uses two sources of real-world messages: (1) a coded (labeled) Twitter dataset composed of messages labeled as cyberbullying or non-cyberbullying and (2) real cyberbullying messages.

Human Evaluation. The human-readable dataset generated by the hybrid social network was then evaluated individually by members of the research team. To evaluate the user records, the team members assessed the user's profile information, stream of messages, and assigned a Bullying Rank between 0 and 100 to reflect the probability the user is experiencing cyberbullying.

Evaluation using the BullyBlocker App. The generated user records were also processed by the BullyBlocker app. To this end, the app was extended by a module that read from the generated dataset instead of obtaining the information from Facebook. This module executed the Bullying Rank Computation task for each user and saved the Bullying Rank values generated by the app.

Evaluation Results

Fig. 1 shows the average error of the Bullying Rank values computed by the proposed BullyBlocker model, with the Bullying Rank (probability that an adolescent is being cyberbullied) expressed as a percentage value (1-100). We compute the error as the absolute value of the difference between the Bullying Rank produced by the app and the average value of the human coding results.

Fig. 2 presents the frequency of errors for various error ranges. The results show that when the weights of Warning Signs and Vulnerability Factors in the BullyBlocker app were set to 50%, the most frequent error values fell within the lowest error ranges.

Fig. 3 shows the distribution of the score difference in the Bullying Rank estimates made by the two annotators who evaluated each case during the Human Evaluation phase.
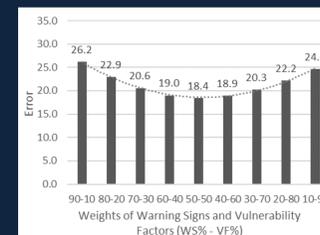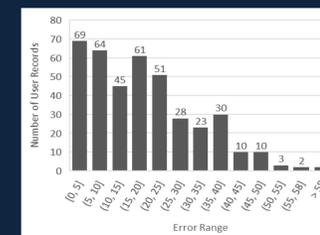

Fig. 1 Average Error for Different Weights
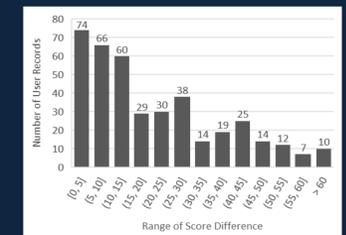

Fig. 2 Histogram of Errors

Fig. 3 Histogram of the Score Difference between Human Coders

Publications

- Y. N. Silva, D. Hall, C. Rich (ASU Student). *BullyBlocker: Toward an Interdisciplinary Approach to Identify Cyberbullying.* Social Network Analysis and Mining (SNAM), 8, 1, 2018.

- Y. N. Silva, C. Rich (ASU Student), J. Chon (ASU Student), L. M. Tsosie (ASU Student). *BullyBlocker: An App to Identify Cyberbullying in Facebook.* The 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 2016.

- Y. N. Silva, C. Rich (ASU Student), D. Hall. *BullyBlocker: Towards the Identification of Cyberbullying in Social Networking Sites.* The 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 2016.

- L. M. Tsosie (ASU Student), Y. N. Silva. *Facebully: Towards the Identification of Cyberbullying in Facebook.* The Grace Hopper Celebration of Women in Computing (GHC), Minnesota, USA, 2013.