# BullyBlocker: Towards the Identification of Cyberbullying in Social Networking Sites*

Yasin N. Silva
Arizona State University
Glendale, AZ, USA
ysilva@asu.edu

Christopher Rich
Arizona State University
Glendale, AZ, USA
cdrich2@asu.edu

Deborah Hall
Arizona State University
Glendale, AZ, USA
d.hall@asu.edu

***Abstract*— Cyberbullying is the deliberate 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. While there have been many studies about the nature and prevalence of cyberbullying, there has been relatively less work in the area of automated identification of cyberbullying in social media sites. The focus of our work is to develop an automated model to identify and measure the degree of cyberbullying in social networking sites, and a Facebook app for parents, built on this model, that notifies them when cyberbullying occurs. This paper describes the challenges associated with building a computer model for cyberbullying identification, presents key results from psychology research that can be used in such a model, describes an initial model and mobile app design for cyberbullying identification, and describes key areas of future work to improve upon the initial model.**

***Keywords—cyberbullying; automated identification; social networks; Facebook***

## I. INTRODUCTION

Over half of adolescents have been bullied online, and about the same number have engaged in cyberbullying; more than one in three young people have experienced cyber-threats online; and well over half of young people do not tell their parents when cyberbullying occurs [1]. Cyberbullying can take multiple forms such as posting hurtful or threatening messages online, spreading rumors on social networking sites, taking and posting unflattering pictures of a person, or circulating sexually suggestive pictures or messages about a person. The consequences of cyberbullying—which can include anxiety, depression, and even suicide—are detrimental on both an individual and societal level.

The goal of our work is to study, design and implement a model to identify cyberbullying in social networking sites. The initial model has been used to build BullyBlocker, a mobile app that identifies cyberbullying on Facebook. BullyBlocker alerts parents of potential cyberbullying instances by providing them with a Bullying Rank that estimates the probability of their child being bullied, allowing them to identify and prevent this form of online aggression. While Facebook is the most common social media platform for teens [2], the principles and design used in BullyBlocker can also be applied to other social networking platforms. Furthermore, similar models could also be used to identify negative outcomes that may result from cyberbullying, such as depression or self-destructive behavior. The key contributions of this paper are:

- The design of an initial model for identifying cyberbullying that builds on previous research findings in the areas of traditional bullying and cyberbullying in adolescents.
- The design and implementation of a mobile app (BullyBlocker) that uses the proposed model to identify cyberbullying in Facebook.
- The identification of challenges and opportunities to integrate the latest results from psychology and social network data analysis to address a problem of great social impact.

## II. BACKGROUND AND RELATED WORK

To develop an effective cyberbullying identification model, we propose building on findings from within the psychology community. There are numerous studies exploring the psychological dimensions of social interactions that can be used to identify the cyberbullying risk factors that a model should consider. Most of the work in identifying bullying among adolescents has focused on traditional bullying, or cyberbullying via mobile or chat-based venues, e.g., [3, 4, 5, 6, 7, 8]. Previous contributions have studied various aspects of bullying and cyberbullying, e.g., whether parents' perception of adolescents' online behavior is causational with adolescents' vulnerability to cyberbullying [3, 4], probabilities of victimization [5] and emotional impact [6, 7] based on age and gender, and measuring the correlation between severity of online aggression and the number of bullies involved [8]. While the results about prevalence and determinants of cyberbullying vary in the psychology literature, there are some important trends and areas of agreement [9, 10] among these results. A key step in our model development process is to identify and use these results to build an automated identification model.

We divided the set of bullying factors into *warning signs* (quantifiable measures like the number of insulting Facebook wall posts) and *states of vulnerability* (circumstances that may increase the probability of experiencing cyberbullying, for example, an adolescent's age and gender). For instance, the survey-driven work in [6, 7] studied the frequency and emotional impact of cyberbullying among groups of adolescents that differed in age and gender. These studies helped us to identify age and gender as two characteristics (states of vulnerability) to consider in the identification of cyberbullying, as well as to estimate the probability of bullying for each group.

* This work was supported by the Dion Initiative for Child Well-Being and Bullying Prevention and ASU NCUIRE awards.

Similarly, the work in [5] concluded that cyberbullying victims are typically adolescents on the "fringe" of various peer groups, e.g., newcomers and members of minority groups. Social media data can be mined to identify if a user belongs to any of these groups. BullyBlocker identifies cyberbullying warning signs and states of vulnerability by (1) analyzing the interaction of an adolescent with his or her network, and (2) obtaining information from the adolescent's Facebook profile.

## III. Automated Identification of Cyberbullying

BullyBlocker analyzes adolescents' interactions with their social network to identify cyberbullying warning signs and states of vulnerability. In the first version of the model we consider an initial set of factors. The main design components of BullyBlocker, our app to identify cyberbullying on Facebook, are presented in Fig. 1. The app is designed for use by the parent or guardian of an adolescent, who will be required to enter the Facebook login information of the adolescent being monitored.

The *Data Collection Module* is the component that uses the adolescent's login information to retrieve all the required information from Facebook, i.e., the list of recent wall posts by other Facebook users, the set of photo comments by other users, user profile information about recently attended schools, etc. The *Cyberbullying Identification Module* then uses the retrieved data to estimate the likelihood that the adolescent is a victim of online aggression. To this end, the application computes a *Bullying Rank* (BR) expression that is based on identified warning signs and states of vulnerability. This rank is used to normalize the intensity of cyberbullying and to simplify the results presented to the parent. Fig. 2 shows that BR is computed based on the values of Warning Signs (WS) and Vulnerability Factors (VF). Each part is given an appropriate initial weight such that the range of BR is [0,100]. The Bullying Rank, together with several aggregated measures such as the number of insulting wall posts, and number of insults in photo comments, is generated by the Cyberbullying Identification Module and loaded in the results page of the BullyBlocker app. This module also generates a set of parent/victim resources, including anti-bullying websites and hotlines.

### A. Measuring Warning Signs

The Warning Signs (WS) component aims to quantify the amount of insulting content received by the monitored adolescent. This component is included to account for the *Group Effect*, as identified in [8], where the number of insults increases the severity of perceived victimization. As shown in Fig. 2, this component is currently computed based on the number of feed (wall) insults and the number of photo insults received by the adolescent during the last $N$ days (currently $N$=60). These raw counts are obtained by running hash-based lookup operations on a table of insults and their variants. The raw insult counts are combined into the *Daily Weighted Insult Count* (DWIC), which currently applies an equal weight to both sub-components and computes the average value per day. The DWIC value is then normalized to be in the [0,1] range. Rather than applying uniform scaling, we use a function that assigns higher weights to initial insults, i.e., after certain large value of daily insults (~30), additional insults tend to have a very minor effect. The function (specified in the WS box in Fig. 2) is plotted in Fig. 3.

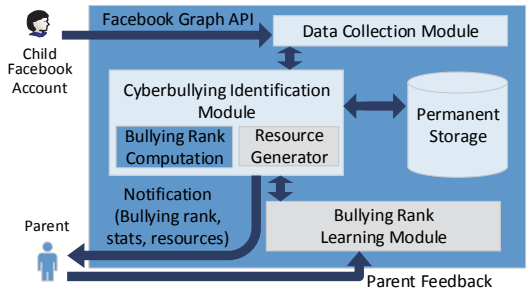


Fig. 1. BullyBlocker Architecture

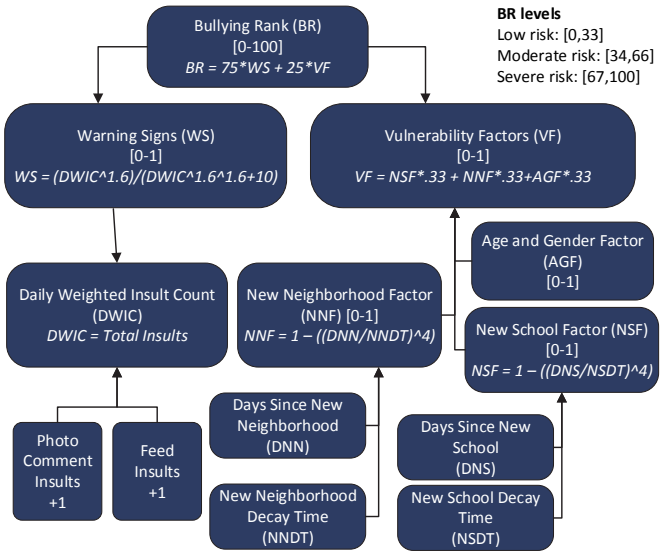


Fig. 2. Bullying Rank Factors

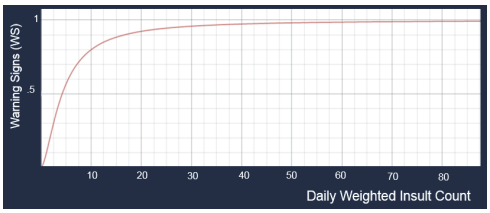


Fig. 3. Warning Signs Vs. Daily Insult Count

### B. Measuring Vulnerability

The Vulnerability Factors (VF) component aims to quantify the level of vulnerability of the monitored adolescent. As shown in Fig. 2, this component is currently computed based on: the Age-Gender Factor, the New Neighborhood Factor, and the New School Factor. Each of these factors is currently weighted equally to compute a VF value in the range of [0,1]. The Age-Gender Factor (AGF) is derived from the statistics of cyberbullying prevalence in different age-gender groups [5]. The New Neighborhood Factor (NNF) and New School Factor (NSF) are introduced to represent the higher vulnerability levels associated with adolescents who recently moved to a new neighborhood and to a new school, respectively. To compute these components, the model uses the number of days since the adolescent moved to a new neighborhood or school. The effect

of these components is assumed to change over time, i.e., the effect should be higher if the adolescent moved recently. To represent this, the model uses a function, specified in the NNF and NSF boxes in Fig. 2, that generates a NNF or NSF value that starts at 1 and decreases over time until it reaches a value of 0.

Fig. 4 shows two screenshots that correspond to the result pages generated for two monitored adolescents.

## IV. Improving the Identification Model

### A. Parent Feedback

One area of future work is the inclusion of mechanisms to improve the accuracy of the initial model. To this end, parents will be able to provide feedback about the app's results. This information paired with the processing logs of the app (e.g., Bullying Rank, sub-components, and aggregated summaries) can be analyzed by domain experts and our team to identify new vulnerability factors and warning signs or modify the weight values and probabilities used in the identification model.

### B. Integrating Machine Learning Components

The cyberbullying identification task can be modeled as a classification problem and multiple machine learning strategies can be considered to implement it. The initial version of the BullyBlocker app could be used to collect a dataset (including parent and domain-expert feedback) that can then be used to train a machine learning based classification model.

### C. Integrating New Vulnerability Factors

There is a rich body of literature in the psychology community that can guide the identification of additional factors. Our hope is that many of these new factors—as was the case with the ones in our initial model—can be integrated into a revised version of the model using data retrieved from social networking sites. For example, we plan to study the integration of the following factors: socio-economic status, race and ethnicity, weight, sexual orientation, and physical disability.

### D. Smart Generation of Parent/Victim Resources

Finding the most relevant resources (e.g., hotlines and anti-bullying organizations) for a specific instance of cyberbullying can be an overwhelming task. The data compiled by BullyBlocker on the specific warning signs and vulnerability factors associated with a cyberbullying instance can be used to generate a customized list of resources.

### E. Synergistic Work with the Psychology Community

The model on which BullyBlocker is based can also help inform the work of psychologists from both a research and clinical practice perspective. The data generated by our identification model can lend further support to previous findings within the psychological literature on cyberbullying and play an instrumental role in devising and testing hypotheses about additional cyberbullying risk factors. Furthermore, aspects of the use of automated tools to identify behavioral issues can also be studied from the psychological perspective. For instance, what level of detail might deter adolescents from providing their parents with their Facebook account login information?

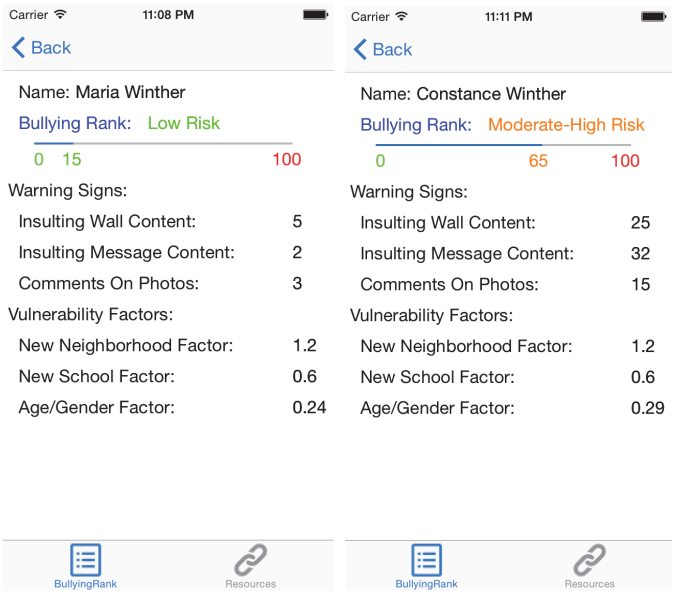


Fig. 4. BullyBlocker Results

## V. Conclusions

This paper proposes a model for cyberbullying identification that builds on the research findings in the psychology community. The paper describes the design of BullyBlocker, an app that implements the proposed model, and presents several ways in which the model can be extended. BullyBlocker aims to have a strong impact on youth in general, by enabling parents to help their children in time to make a difference.

## References


[1] http://www.bullyingstatistics.org/content/cyber-bullying-statistics.html.

[2] http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015/.

[3] R. P. Ang, W. H. Chong, S. Chye, and V. S. Huan. Loneliness and generalized problematic Internet use: Parents' perceived knowledge of adolescents' online activities as a moderator. Computers in Human Behavior, 28 (4), 1342-1347.

[4] D. M. Law, J. D. Shapka, and B. F. Olson. To control or not to control? Parenting behaviours and adolescent online aggression. Computers in Human Behavior, 26 (6), 1651-1656.

[5] J. Piazza and J. M. Bering. Evolutionary cyber-psychology: Applying an evolutionary framework to Internet behavior. Computers in Human Behavior, 25 (6), 1258-1269.

[6] R. Ortega, P. Elipe, J. A. Mora-Merchin, J. Calmaestra, and E. Vega. The emotional impact on victims of traditional bullying and cyberbullying: A study of Spanish adolescents. Journal of Psychology, 217 (4), 197-204.

[7] T. E. Waasdorp and C. P. Bradshaw. Examining student responses to frequent bullying: A latent class approach. Journal of Psychology, 103 (2), 336-352.

[8] J. J. Dooley, J. Pyżalski, and D. Cross. Cyberbullying versus face-to-face bullying: A theoretical and conceptual review. Journal of Psychology, 217 (4), 182-188.

[9] D. Wolke, T. Lereya, and N. Tippett. Individual and Social Determinants of Bullying and Cyberbullying. Cyberbullying. Ed. T. Vollink, Ed. F. Dehue, Ed. C. Guckin. Routledge, 2016. 26-53.

[10] J. W. Patchin and S. Hinduja. Cyberbullying - An Update and Synthesys of the Research. Cyberbullying Prevention and Response. Ed. J. W. Patchin, Ed. S. Hindija. Routledge, 2012. 13-35.