BullyBlocker: An App to Identify Cyberbullying in Facebook*

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

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

Jaime Chon
Arizona State University
Glendale, Arizona, USA
jchon@asu.edu

Lisa M. Tsosie
Arizona State University
Glendale, Arizona, USA
lmtsosi1@asu.edu

***Abstract*— Cyberbullying is the most common online risk for adolescents, and it has been reported that over half of young people do not tell their parents when it occurs. Cyberbullying involves the deliberate use of online digital media to communicate false or embarrassing information about another person. While previous work has extensively analyzed the nature and prevalence of cyberbullying, there has been significantly less work in the area of automated identification of cyberbullying, particularly in social networking sites. The focus of our work is to develop a computational model to identify and measure the intensity of cyberbullying in social networking sites. In this paper, we present and demonstrate BullyBlocker, an app that identifies instances of cyberbullying in Facebook and notifies parents when it occurs. This paper presents the most relevant characteristics of our initial cyberbullying identification model, key app design and implementation details, the demonstration scenarios, and several areas of future work to improve upon the initial model.**

Keywords—cyberbullying; automated identification; social networks; Facebook

I. INTRODUCTION

Cyberbullying is a critical problem that many adolescents experience. It can take multiple forms such as posting hurtful or threatening messages online, taking and posting unflattering pictures of a person, spreading rumors or false information on social networking sites, or circulating sexually suggestive pictures or messages about a person. Over half of adolescents have been bullied online, and about the same number have engaged in cyberbullying. Furthermore, 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 they experience cyberbullying [1]. Cyberbullying has several detrimental consequences at the individual and societal level including anxiety, depression, and even suicide.

While previous work in the area of psychology has addressed the nature and prevalence of cyberbullying, e.g., cyberbullying measures for mobile and chat-based venues, there has been relatively less work on the actual implementation of automated models or applications to identify cyberbullying.

The goal of the work reported in this paper is to study, design and implement a model to identify cyberbullying in social networking sites. This model is the core component of BullyBlocker, an app aimed at identifying instances of cyberbullying in Facebook. BullyBlocker is an app built to be used by parents of adolescents. After a parent enters the login information of the adolescent he or she wants to monitor, the app analyzes the interaction of the adolescent with his or her social network and notifies the parent when cyberbullying is detected. BullyBlocker generates a key value — the Bullying Rank — that estimates the probability of the adolescent being bullied. Facebook was selected as the initial social networking site because it is the most common social media platform for teens [2]. However, the principles and design guidelines used in BullyBlocker can also be applied to other social networking platforms like Instagram and Twitter. Furthermore, a similar model could be applied to identify depression or self-destructive tendencies.

In this paper, we present and demonstrate the first version of BullyBlocker, an app designed for parents that monitors the Facebook interactions of their adolescents and notifies them when cyberbullying is detected. The paper includes the description of: (1) the cyberbullying identification model's primary characteristics, (2) design and implementation details of the app, (3) several demonstration scenarios that will enable attendees to get a hands-on experience with the app, (4) and several ways in which the initial model and app can be improved and extended (we expect to engage in interesting conversations with other researchers about these ideas).

The remainder of this paper is organized as follows. Section II presents the background and related work, Section III describes the proposed model and app design guidelines, Section IV describes the demonstration scenarios, Section V presents several areas of future work, and Section VI concludes the paper.

II. BACKGROUND AND RELATED WORK

One of the key characteristics of the approach we followed to develop the cyberbullying identification model was to consider previous results in the area of psychology. The research community in this area has produced many studies that can be used to identify possible risk factors to include in an automated identification model.

A good number of previous papers in this area focus on the study of traditional bullying and cyberbullying via mobile or chat-based venues, e.g., [3, 4, 5, 6, 7, 8]. These studies have considered various aspects associated with traditional bullying and cyberbullying, e.g., whether parental perception of adolescents' online behavior is causational with adolescents' vulnerability to cyberbullying [3, 4], measuring the severity of online aggression in correlation to the number of bullies

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

involved [8], and probabilities of victimization [5] and emotional impact [6, 7] based on age and gender. While the specific results about the prevalence and determinants of cyberbullying vary among different studies in the psychology literature, several recent studies have identified important commonalities and trends among these results [9, 10]. The identification and use of these common results to design an initial cyberbullying identification model was an important step in our model development process. Many papers have been published in this area. Several of them have been surveyed in [9] and [10]. We mention next some specific papers whose results have been used in the first version of our model.

In our design, we divided the set of cyberbullying risk factors into two broader groups: *warning signs* and *states of vulnerability*. Warning signs are quantifiable measures like the number of insulting posts or the number of offensive comments in a picture. States of vulnerability are the circumstances that could increase the probability of experiencing cyberbullying such as the minor's age or gender. For instance, the survey-driven work in [6, 7] studied the frequency and emotional impact of cyberbullying in different age-gender groups. The results of these studies were used to identify gender and age as two potential states of vulnerability as well as specific probability values of cyberbullying risk for each age-gender group. Likewise, one of the key findings in [5] was that cyberbullying victims are typically adolescents on the "fringe" of various peer groups, e.g., newcomers, members of minority groups (Hispanics, African American, Indian American, etc.), and people with mental or physical disabilities. The data available in social networking sites like Facebook can be mined to identify if an adolescent belongs to any of these groups.

Currently, the BullyBlocker app identifies cyberbullying warning signs and states of vulnerability by (1) analyzing the interaction of a minor with his or her network through wall posts, picture comments, and messages, and (2) obtaining information from the minor's Facebook profile, like home address and schools attended. The specific way in which the identified measures are processed in our model is described in Section III. In the future, we also plan to integrate the parent as a source of information. The parent can provide key information about the behavior and personality of the minor as well as provide information missing in the minor's Facebook profile, e.g., the date of the minor's move to a new neighborhood.

Preliminary ideas of the BullyBlocker model have been published in two posters [11, 12].

III. Automated Identification of Cyberbullying

BullyBlocker, our app to identify cyberbullying in Facebook, analyzes the interactions of adolescents with their social networks to identify cyberbullying warning signs and states of vulnerability. The first version of the identification model considers an initial set of risk factors. The supported warning signs include: the number of insulting wall posts and the number of embarrassing comments on photos. The supported states of vulnerability include: age, gender, and peer social status (being new to a neighborhood or school).

The architecture of the BullyBlocker app, which includes the core app design components, is presented in Fig. 1. The app is

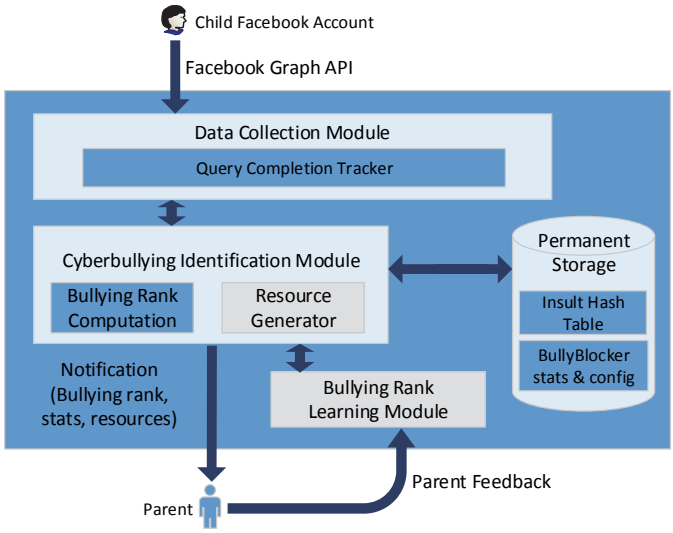

Fig. 1. BullyBlocker Architecture

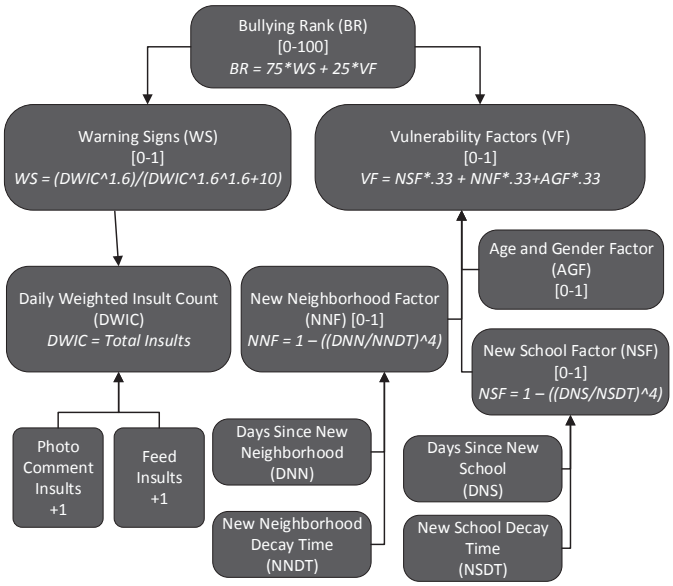

Fig. 2. Bullying Rank Factors

aimed to be used by parents or guardians who want to monitor their adolescents. The parent is required to enter the Facebook login information of his or her minor. This login information is used by the app to retrieve the relevant data from Facebook.

The Data Collection Module is the component that collects all the relevant 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 data of wall posts and picture comments is provided by Facebook using a multilevel chain of pages. For instance, a root page can contain a set of wall posts, and a first level page can contain comments on the root level posts. Facebook currently supports two levels of comments. The app submits asynchronous queries to retrieve different pages. To properly track the completion of these queries, the app uses a

query tracking mechanism (tracking tree structure) that records the state of each submitted query. Also, to comply with the query rate limit imposed by Facebook, the app adjusts the frequency of queries per second to comply with the required threshold.

The Cyberbullying Identification Module is in charge of using the retrieved data to determine if the minor is experiencing cyberbullying. The core output of this component is the Bullying Rank (BR), which is computed based on the identified warning signs and states of vulnerability. This value represents the probability of experiencing cyberbullying and simplifies the presentation of the results to the parent(s). Fig. 2 shows the various components used to compute the Bullying Rank. Observe that the Bullying Rank (BR) value is computed based on the values of Warning Signs (WS) and Vulnerability Factors (VF). Each subcomponent is given an appropriate weight such that the range of BR is [0,100]. For reporting purposes, we divided this range into three intervals: low risk [0,33], moderate risk [34,66], and severe risk [67,100].

The Bullying Rank and several key aggregated measures, e.g., the number of feed (wall) insults, the number of insults in photo comments, etc., are outputted by the Cyberbullying Identification Module and presented to the parent in the results page of the BullyBlocker app. This module also generates a set of anti-bullying resources, e.g., websites and hotlines, that can be used by parents to learn more about cyberbullying and to find information about ways to deal with it. Some of the information generated by this module is stored in the mobile device's permanent storage. The recorded information includes the Bullying Rank and its various components, the most recent dates in which the minor moved to a new neighborhood or school, etc.

Modules that will be the focus of our future work include the Bullying Rank Learning Module and the Resource Generator. Section V elaborates on these components.

A. *Measuring Warning Signs*

The goal of the Warning Signs (WS) component is to quantify the amount of insulting content received by the monitored adolescent. This component accounts for the *Group Effect* (identified in [8]) where higher number of insults are associated with increased severity of perceived victimization. As stated in Fig. 2, this component is computed based on the number of feed (wall) insults and the number of photo insults received by the minor during the most recent *N* days (*N* is a parameter currently set to 90). To identify individual insults, the app searches for insults (and possible variants) in the text of feed posts and photo comments. This task performs hash-based lookup operations on a table of insults and their variants. The initial insult counts are combined into the Daily Weighted Insult Count (DWIC), which applies an equal weight to both sub-components and computes the weighted number of insults per day. Then, the DWIC value is scaled to be in the [0,1] range. Rather than applying uniform scaling, the model uses an approach that considers that initial insults up to certain level (about 30 insults per day) should have greater effect than insults after this threshold. The scaling function (specified in the WS box in Fig. 2) is plotted in Fig. 3. As shown in this figure, going from 0 to 10 daily insults has a greater effect in the function value than going from 70 to 80 insults.

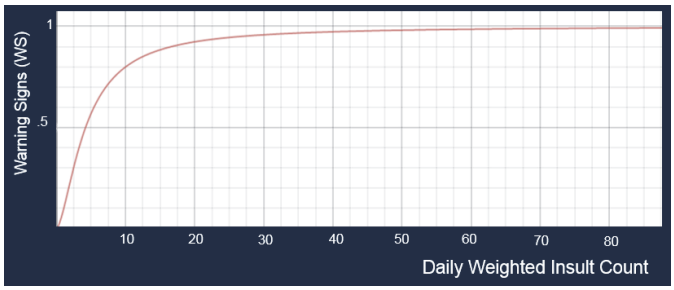

Fig. 3. Warning Signs Vs. Daily Insult Count

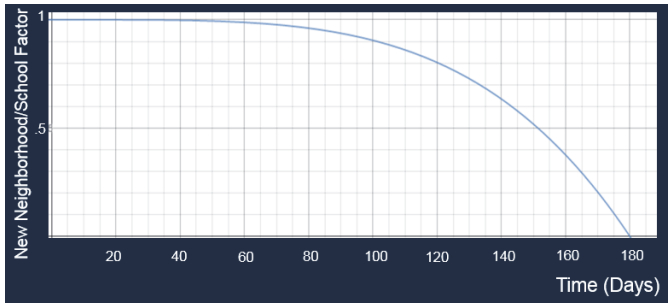

Fig. 4. New Neighborhood/School Factor Vs. Time

Important open questions about measuring warning signs are:

- Should certain types of insult be weighted more heavily than others? For instance, should photo insults have a higher weight than wall insults?
- How can measures like the number of people who *liked* a given insulting post be incorporated in the model?
- Currently, we consider the insults received in the last 90 days (N=90). What is the optimal value of N? Should recent insults bare more weight than older ones?

B. *Measuring Vulnerability*

The goal of the Vulnerability Factors (VF) component is to measure how vulnerable a minor is to be a cyberbullying victim. As specified in Fig. 2, this component is computed using the Age-Gender Factor, the New Neighborhood Factor, and the New School Factor. In the current model these subcomponents are equally weighted and generate a VF value in the [0,1] range.

The Age-Gender Factor (AGF) is determined using the statistics of cyberbullying prevalence based on age and gender [5]. The New Neighborhood Factor (NNF) and New School Factor (NSF) increase the vulnerability level if a minor has recently moved to a new neighborhood or to a new school, respectively. The effect of NNF and NSF is considered to be dependent on the number of days since the minor moved to a new neighborhood or school. This is represented in the model using the function specified in the NNF and NSF boxes in Fig. 2 and plotted in Fig. 4. This function generates a NNF or NSF value in the range of [0,1]. The value of 1 is generated when the minor moves to a new neighborhood or school, and then it decreases over time. The functions produce a value of 0 when the number of days is equal to a system parameter (*New*

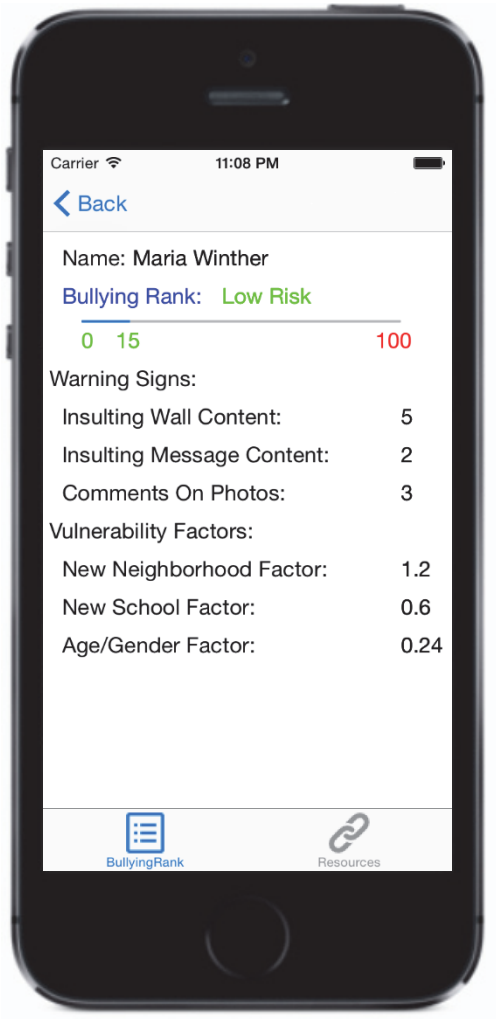

Fig. 5. BullyBlocker Results – Low Risk

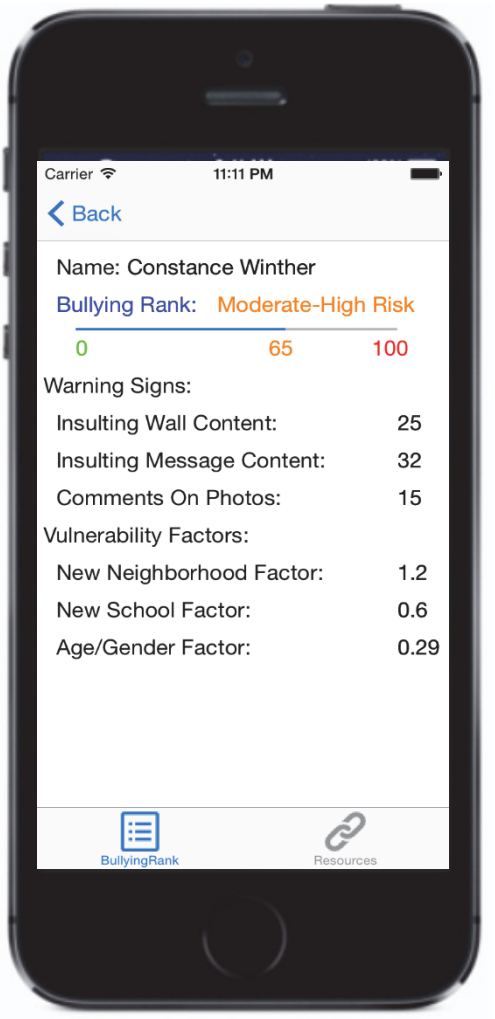

Fig. 6. BullyBlocker Results – Moderate-High Risk

Neighborhood Decay Time or *New School Decay Time*) currently set to 180.

Important open questions about measuring vulnerability are:

- Should certain types of vulnerability factors be given greater weight than others? For instance, should the Age-Gender Factor have a higher weight than the New School Factor?
- How can we measure the contribution of positive social interactions in relation to the model? How should the number of new friends made since moving to a new school or neighborhood be weighted?
- Currently, the decay time parameters are set to 180. What is the optimal value of this parameter?

IV. Demonstration Scenarios

The goals of this demonstration are (1) to provide attendees with hands-on experience with the BullyBlocker app under several scenarios, and (2) to establish a technical dialogue with other researchers about mechanisms to improve and extend the initial BullyBlocker model and app.

To provide a hands-on experience with BullyBlocker, we will install the app on several devices and use multiple real and artificial Facebook users to show the output generated by the app (Facebook allows the creation of artificial users for app development and testing).

We will include scenarios with various levels of bullying probability. For example, figures 5 and 6 show two screenshots that correspond to the result pages generated for two monitored minors, Maria and Constance Winther. While the Bullying Rank of Maria is relatively low (15), the one for Constance is significantly higher (65) due to a much higher value of Warning Signs.

During the demonstration session, we will show not only the output directly accessible through the app, but also the breakdown of the BR factors, which allows the attendee to compare the contributions and dissect the significance of the factors towards the rank. Attendees will be able to use the app and test it using multiple users.

The demonstration session will also be a great opportunity to engage in conversations with other researchers about how the proposed model can be improved and extended. Particularly, we plan to interact with experts in Psychology and Machine

Learning to discuss the integration of additional warning signs and states of vulnerability. Several topics and ideas for this segment are presented in Section V.

Additional project details and progress updates are available in the BullyBlocker project website [13].

V. Improving the Identification Model

A. Parent Feedback

One possible extension of the model is to integrate parent feedback. To this end, parents who use the application will be able to provide feedback about the accuracy of the model. This information, paired with the processing logs of the BullyBlocker app, can be analyzed by domain experts and our team to improve the accuracy, e.g., by identifying new vulnerability factors and warning signs or modifying the weight values and probabilities used in the identification model.

B. Integrating Machine Learning Components

Another possible improvement is the integration of machine learning techniques. The cyberbullying identification task can be modeled as a classification problem and multiple strategies can be used to implement it, e.g., Support Vector Machines and Naïve Bayes. Given that classification is a supervised learning task, and thus requires a training dataset, we can use the initial version of the app to collect a training dataset (including parent and expert feedback). The machine learning approach should also integrate the key findings about prevalence and determinants of cyberbullying identified previously in the psychology literature.

C. Integrating New Vulnerability Factors

One of the most interesting areas of future work is the integration of new vulnerability factors into the cyberbullying identification model. There is a rich body of literature in the psychology community that can guide the identification of appropriate factors. We plan to investigate the addition of the following factors: socio-economic status, race and ethnicity, weight, sexual orientation, physical and mental disability, etc.

D. Smart Generation of Parent/Victim Resources

One of the challenges that parents face after they find out that their minor is experiencing cyberbullying is finding the proper resources (e.g., anti-bullying organizations, hotlines, etc.). Since BullyBlocker is aware of the specific warning signs and vulnerability factors, it can use this information to generate a customized list of relevant resources.

E. Synergistic Work with the Psychology Community

The work in models like BullyBlocker can also benefit the work in the psychology community. Automated identification models can be tools to verify previous results, and provide data to identify and test hypotheses about additional cyberbullying 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 about identified cyberbullying instances are parents comfortable receiving through the app? What about the victims? Should the app report only aggregated information or specific insults?

VI. Conclusions

Cyberbullying is the most common online risk for adolescents. However, there has not been much work on its automated identification in social networking sites. This paper presents a model and app (BullyBlocker) to identify cyberbullying instances in Facebook. The paper presents the design elements of BullyBlocker, describes hands-on demonstration scenarios, and discusses areas of future work.

Acknowledgment

The authors would like to thank ASU students Tara Tucker, Jason Cheney, Gohar Hunter, Aakanxu Shah, and Mohan Thorat for their contributions to the implementation of the BullyBlocker app.

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.

[11] L. M. Tsosie and Y. N. Silva. Facebully: Towards the Identification of Cyberbullying in Facebook. The Grace Hopper Celebration of Women in Computing (GHC), Minnesota, USA, 2013.

[12] Y. N. Silva, C. Rich, and D. Hall. BullyBlocker: Towards the Identification of Cyberbullying in Social Networking Sites. The IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), California, USA, 2016.

[13] http://www.public.asu.edu/~ynsilva/BullyBlocker/.