BullyAlert- A Mobile Application for Adaptive Cyberbullying Detection

Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra

University of Colorado Boulder, CO, USA
{rahat.rafiq, richard.han, qin.lv, shivakaht.mishra}@colorado.edu

Abstract. Due to the prevalence and severe consequences of cyberbullying, numerous research works have focused on mining and analyzing social network data to understand cyberbullying behavior and then using the gathered insights to develop accurate classifiers to detect cyberbullying. Some recent works have been proposed to leverage the detection classifiers in a centralized cyberbullying detection system and send notifications to the concerned authority whenever a person is perceived to be victimized. However, two concerns limit the effectiveness of a centralized cyberbullying detection system. First, a centralized detection system gives a uniform severity level of alerts to everyone, even though individual guardians might have different tolerance levels when it comes to what constitutes cyberbullying. Second, the volume of data being generated by old and new social media makes it computationally prohibitive for a centralized cyberbullying detection system to be a viable solution. In this work, we propose BullyAlert, an android mobile application for guardians that allows the computations to be delegated to the hand-held devices. In addition to that, we incorporate an adaptive classification mechanism to accommodate the dynamic tolerance level of guardians when receiving cyberbullying alerts. Finally, we include a preliminary user analysis of guardians and monitored users using the data collected from BullyAlert usage.

Keywords: Cyberbullying · Mobile Application · Detection

1 Introduction

Cyberbullying in Online Social Networks (OSNs) has seen an unprecedented rise in recent years. Continued democratization of internet-usage, advancements, and innovations in the area of online social networking and lack of diligent steps to mitigate the effect of cyberbullying by the concerned entities, all have contributed to this unfortunate consequence. The constant threat of cyberbullying in these OSNs has become so expansive and pervasive that it has been reported that in America alone, more than fifty percent of teenage OSNs users have been affected by the threat of cyberbullying [4]. While real-life bullying may involve verbal and/or physical assault, cyberbullying is different in the sense that it occurs under the umbrella of an electronic context that is available 24/7, thereby

rendering the victims vulnerable to its threats on a constant and relentless basis. This unique feature of cyberbullying subjects the victims to devastating psychological effects that later cause nervous breakdowns, low self-esteem, self-harm, clinical depression, and in some extreme cases, suicides [10],[8]. Disturbing events of teens committing suicides after being victimized by cyberbullies have also been reported [9], [30]. Moreover, nine suicide cases have already been attributed to cyberbullying in Ask.fm alone [2]. Cyberbullying has been reported as one of the many potential factors, if not the only factor, for these suicides [3]. Figure 1 shows an example of a cyberbullying instance on Instagram.

Fig. 1. An example of cyberbullying on Instagram.

There have been numerous works on cyberaggression and cyberbullying by performing thorough analyses of the labeled media sessions from a social network. Cyberaggression is defined as a type of behavior in an electronic context that is meant to intentionally harm another person [16]. Cyberbullying is defined in a stronger and more specific way as an aggressive behavior that is *carried out repeatedly* in OSNs against a person who *cannot easily defend himself or herself*, creating a power imbalance [16],[22],[13],[16],[21],[29]. Developing classifiers for cyberbullying detection for different social networks such as Twitter, Ask.fm, Vine, Instagram, YouTube etc have been proposed as part of most of the previous works[28],[11], [15], [35], [6], [26], [27]. System-related challenges such as responsiveness and scalability of a potential cyberbullying detection system have also been proposed [25]. However, there are two key issues that a potential centralized cyberbullying system faces:

– A centralized cyberbullying detection system still needs to have a lot of computing resources to be able to accommodate all the OSNs in the world. Therefore, a system solution that can accommodate all the OSNs currently avail-

able will be faced with a daunting challenge of meeting the computational-resource requirements to maintain an acceptable responsiveness
- A potential centralized cyberbullying detection system, such as one presented in [25], makes use of a cyberbullying classifier component that is applied generally to all the guardians. We argue that different guardians will have different tolerance levels, which in turn, might be dependent on their personal preferences, demographic information, location, age, gender, and so on. So, we argue that a system solution that allows different levels of cyberbullying alerts to be sent to parents based on their tolerance levels is the most natural solution.

In this paper, we present the design and implementation of an Android mobile application for guardians: BullyAlert. This mobile application allows the guardians to monitor the online social network activities (currently only supports Instagram) of their kids and get notifications whenever the monitored social network profiles receive a potential cyberbullying instance. The reasons for developing this mobile application are twofold. First, it allows us to delegate classifier computations of cyberbullying detection to the hand-held devices of the guardians, thereby reducing the computational resources needed for a potential centralized cyberbullying detection system. Second, BullyAlert allows the guardians to give the resident classifier feedback about how right or wrong each notification is. The resident classifier then updates itself accordingly to calibrate its tolerance level with that of the guardian using it. This mechanism allows for personalized cyberbullying notification of an individual guardian.

We make the following contributions in this paper.

- We propose the design and implementation of an android application, BullyAlert.
- We present a preliminary user-experience analysis of the guardians who downloaded the mobile application by using the current crop of data
- We present a preliminary comparison the behavior of the users who were being monitored by the guardians with the general population of Instagram to derive some initial key insights by leveraging the current collection of data

2 Related Works

As mentioned earlier, the majority of researches on cyberbullying have focused on improving the accuracy of detection classifiers. Collection, mining and analyses of different social media networks such as Twitter [28], Ask.fm [18], YouTube[5], Instagram [12],[6], chat-services [15], Vine [26] have been performed for cyberbullying, cyberaggression[34], harassment [35], aggression under the hood of anonymity [18] and predatory behaviors [14]. As a natural next step, many works have proposed accurate classifiers to detect and predict cyberbullying incidents as well as finding most contributing factors to successful classification of cyberbullying incidents [20][6].

The scalability and responsiveness challenges of a potential cyberbullying detection system have also been put under the microscope recently [25]. Several

systems have been proposed with scalable architectures to detect cyberbullying in social media networks [35], [33], [7]. However, none of the previous researches addressed the issue of the individual tolerance level of the guardians when it comes to the detection of cyberbullying, which might be a factor dependent on many variables such as the age of the person being monitored, demographics information, gender, and so on. Moreover, due to many novel social networks being introduced every year[32] and the enormous amount of data being generated by the existing ones such as Instagram[17] and YouTube[19], we argue, eventually it will be computationally prohibitive to sustain a centralized cyberbullying detection architecture. Some mobile applications are available for guardians, however, they are either mostly outgoing packet network sniffing tools[7] or profanity detection apps[31]. None of them incorporates a sophisticated cyberbullying classification or a mechanism to enable dynamic tolerance-level detection.

3 System Design and Implementation

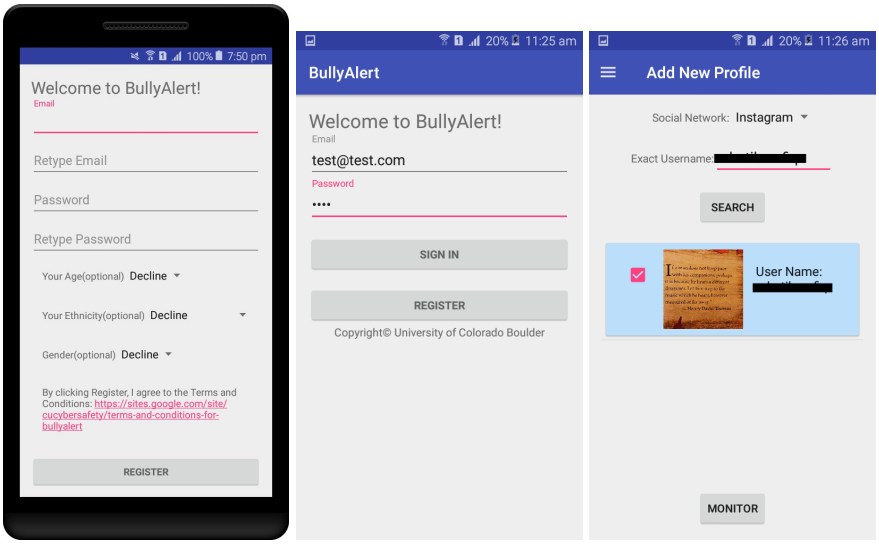
This section presents the design, implementation, and architecture of BullyAlert. We begin by describing the typical user work-flow through a series of use cases, and then present the architecture and implementation of different components of BullyAlert.

3.1 Use Cases

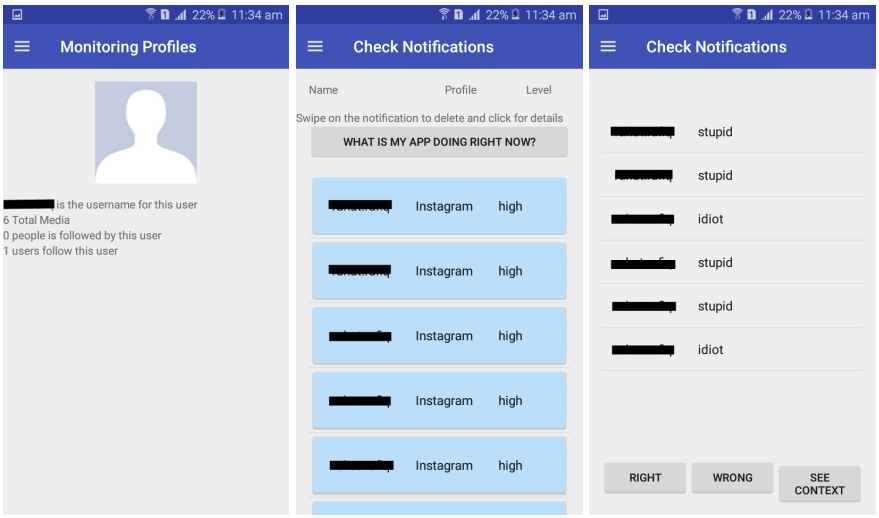
Guardian registers After a guardian downloads the application and opens it, the screen in Table 1a is presented. The Guardian has to enter unique email id and password to be able to register into our system. Options to divulge additional information such as age group, gender, and ethnicity, are also provided. Each of these information is presented through a drop-down list. The Guardian can also use the option to decline to give this information.

Guardian logins The login screen is shown in Table 1b has two input fields. These fields ask for the email and password used to register into the system. After clicking the log-in button, the guardian is directed to the dashboard of the application.

Guardian Searches for users to monitor The Guardian can search for public Instagram profiles by going to the user search component shown in Table 1c. At first, the guardian selects the social network profile from the drop-down list. *Right now, we are only supporting public Instagram social network profiles.* Then in the text field, the username of the user to be monitored is typed. When the search button is clicked, a list of users matching the username entered is shown along with the associated profile pictures for the guardian to facilitate a better identification. To start monitoring a profile, the guardian has to select the profile and then click the monitor button.



(a) Register (b) Log In (c) User Search



(d) User Details (e) Notification List (f) Notification Details

Table 1. BullyAlert Application

Guardian examines user profile information details The Guardian can see the basic profile details of the users being monitored, as shown in Table 1d. This page shows guardians the current profile picture, the number of total media shared, and the number of total followers and followings of the user being monitored.

Guardian gets a list of notifications Table 1e shows the screen that the guardian sees when a host of notifications are present in the dashboard. The list has three columns, the first column shows the username of the profile where this cyberbullying notification has originated, the second column indicates the social network and finally, the third column outlines the application's classifier's perceived level of severity for this notification. Currently, the application has two levels of severity, namely low and high. The Guardian can see the details of the notification by clicking the individual notification boxes.

Guardian examines notification details and give feedback To enable the guardians to see the full context of a particular notification and give feedback as to whether the application right or wrong in terms of the severity level, table 1f is presented. The Guardian has options to click the "see full context" button which will then load not just the latest comments but also the previous comments of that media session. This enables the guardians to get a full picture of the happenings in the media session. The guardian can give feedback through the two buttons, namely right or wrong. This feedback is then used by the application to calibrate its tolerance level according to that of the guardian.

3.2 Architecture and Implementation

This section describes the architecture and implementation of the BullyAlert application's different components. Figure 2 presents the architecture diagram of the BullyAlert system. The guardian communicates with the BullyAlert application for registering, logging in, and getting notifications for potential cyberbullying instances. The application sends guardian data, notification data, and feedback data to the BullyAlert server. The application also contacts the BullyAlert server for authenticating a user log-in. Moreover, the application implements a polling mechanism by which it periodically collects media session data of the Instagram-users (who are being monitored by the guardians) from the Instagram servers.

BullyAlert Server BullyAlert server is responsible for the following:

- During the registration process, it is responsible for checking that the registration information is verified. It first checks if the email that is being used to register is unique in the system and the password is at least six characters. When the registration is successful, the server stores both the username and

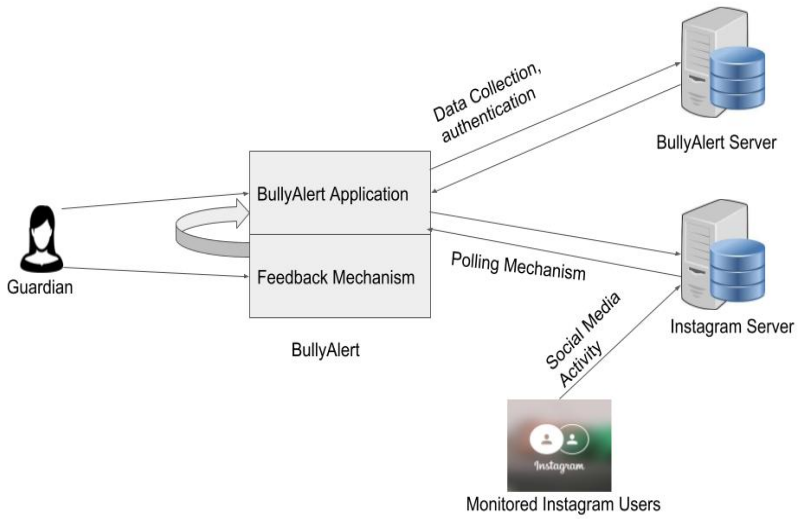


Fig. 2. BullyAlert Architecture

the password in an encrypted format. In addition to this, the server stores the demographic information provided by the guardian through the registration form, such as age group, gender, and ethnicity.

- During login, the server is responsible to check the login credentials of a guardian with the system's stored credentials. It gives an error if the login credentials are not verified which is then in turn shown to the user.
- Storing all the notifications of the guardians. Every time a guardian receives a notification, the application sends the notification meta-data to the server. The data consists of a list of comments for which the notification was raised, the user-name in whose profile the notification was raised, and the severity level of the notification (high or low).
- Storing the feedback that a guardian gives to the application for a particular notification. Every time the guardian gives the application feedback (right or wrong), the application sends the relevant data (list of comments, the application's perceived severity level, the guardian's feedback) to the server.
- Storing each guardian's resident classifier information. This is to facilitate continuity of the guardian's classifier so that when the guardian uninstalls the application, the classifier will keep being stored in our server. This means that, if the guardian, at some later time, chooses to re-install the application, the old classifier will become the classifier of the application instead of the general one.

We have used MongoDB, RESTful API, and node.js for the implementation of this component. The code for this can be found in [24].

Adaptive Classifier An adaptive classifier for each guardian is more suitable for our application than a general classifier for every guardian because of the potentially subjective nature of cyberbullying. We hypothesize that each guardian will have their tolerance level when it comes to cyberbullying, which in turn, can be dependent on several factors, such as gender, age, race, etc.

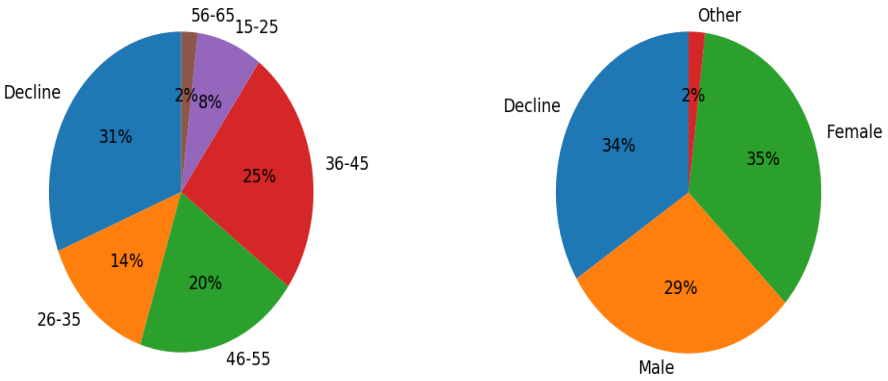
The ways we develop this adaptive classifier are as follows. First, we incorporate a feedback mechanism in our application by which, the guardians, upon receiving a potential cyberbullying notification, will be able to give us feedback saying how right or wrong the notification is. *We also show the guardians a list of other media sessions which were not deemed as bullying by our application, to make sure we also get feedback for media sessions which were not in the potential bullying notifications page. This is to enable the application to keep track of the false negatives in addition to false positives.* Second, we use the logistic regression classifier from [25] for the implementation of the application's resident cyberbullying detection component. Every time an instance of the classifier gets feedback, the feedback data encapsulate the media session's list of comments for which the alert was raised and the guardian's label (right or wrong). This datum is considered as a labeled training data for the resident classifier. Feature values described for the logistic regression classifier used in [25] are extracted for each of these feedback sessions. Upon converting the feedback data into training data, we then perform stochastic gradient descent [23] for the resident classifier. Each parent's classifier then reaches a different local optimum, thereby facilitating the adaptive nature of the classifier.

For the guardians whose numbers of feedback are not substantial enough to perform an individual adaptation process, we implement the following. We first collect all the feedback given by all the guardians in our server. Then, based on all this feedback, we update our general classifier that was used by the guardians when they first install our application. We call this *updated general classifier*. This updated general classifier is then propagated to the guardians who don't have enough individual feedback to make sure their classifiers are updated as well. The implementation code can be found in [1].

Polling Mechanism The polling mechanism is responsible for the following:

- When the guardian searches for a particular user by username, this mechanism fetches the user profiles of which the username-string is a match.
- After a monitoring request of a user by a guardian is approved, the polling mechanism starts polling that user profile every hour for any new posts. This is to make sure the app is updated with the latest media postings of the monitored user.
- In addition to polling for newly posted media, this component is also responsible for getting the newest comments for all the media posted by the user. Every time a host of new comments is posted for a media session, this mechanism fetches those new comments and sends this newest media session data to the adaptive classifier component for classification.

4 User Data Analysis



(a) age group (b) gender

Table 2. BullyAlert Guardian's Gender and Age Distribution

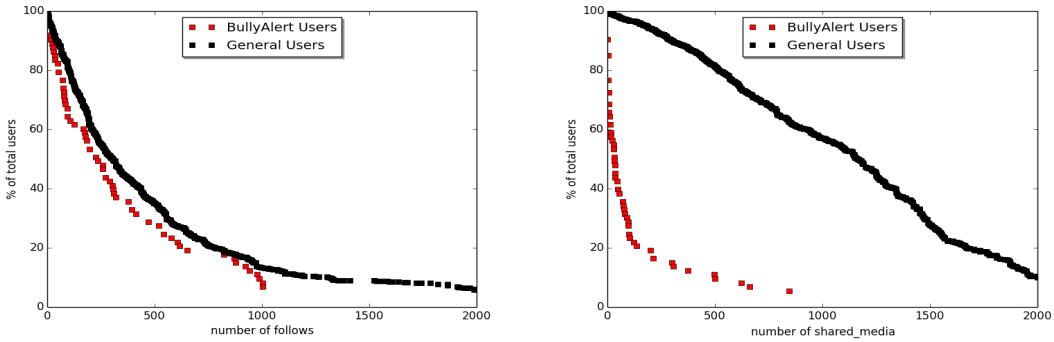
This section presents a preliminary analysis of the data collected until now from BullyAlert. First, it explores the guardian data and then it performs a comparison of social network behaviors between the general Instagram population [18] and the users who were being monitored by the guardians who downloaded our application.

When a guardian registers into our system, in addition to the email and password, we also ask them to provide their gender, ethnicity and age information, if they choose to divulge those. Table 2a and 2b show the distribution of gender and age-group of the 100 guardians who have downloaded BullyAlert until now. From the distributions, it is fairly clear that a substantial portion of the people chose not to provide the demographic information, 31 and 34 percent for age group and gender respectively. In addition to that, the most prominent age group and gender where 36-45 and female respectively.

The reasons we collected this demographic information are twofold. First, when we start to get the classifier feedback data for all the guardians for their different tolerance levels, we will want to investigate if there are any correlations between different guardians' tolerance levels and their demographic information. Second, if we do find that people with the same demographics tend to have the same tolerance levels, we will then want to be able to build different general classifiers for different clusters where each cluster hosts guardians with similar demographics. *While we acknowledge that these 100 guardians' data is an insufficient representation of guardians, we postulate that this preliminary*

demographics distribution still introduces us to a new systems challenge: what about guardians who do not provide demographic information and thus will not belong to any particular cluster by default?

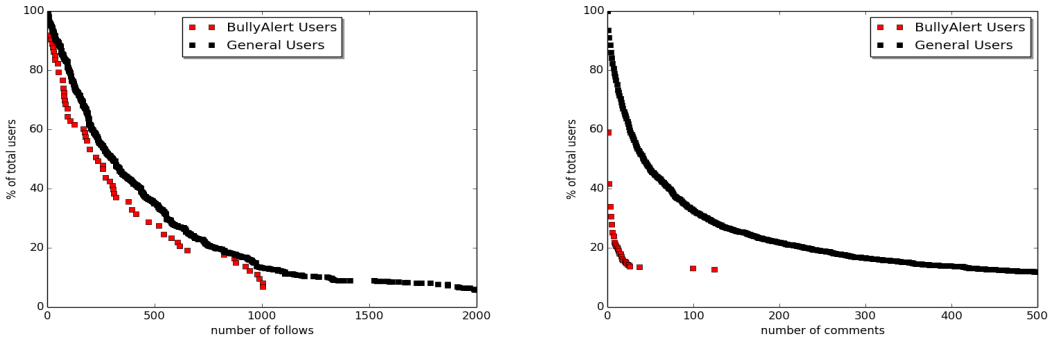
Next, we investigate a comparison analysis between the Instagram users who are being monitored by our application and the general Instagram population, a data-set collected from [18]. First, we compare both sets of users' follow and media-sharing activities. Table 3a and 3b show the CCDF of both set of users' number of people they follow and number of medias they have shared in their profile. It can be seen that the follows activity of both sets follow the same pattern, which is understandable because a user can only follow so many people. But there is a discernible difference in the media sharing activity. The general population's line tends to fall far slower in the graph than that of the BullyAlert-monitored users, with almost 80 percent of the BullyAlert users having less than 100 shared media in their profile. This means that the users who are monitored by the guardians tend to be not as active as the general population. This particular observation also poses an interesting system perspective. Because most of the people who are likely to be monitored by our application will not be sharing as much media, *we can afford to incorporate some sophisticated machine learning classifiers in the application instead of worrying about responsiveness, discussed in [25].* Again, we like to emphasize here that these are preliminary derivations drawn from our current small set of collected data.



(a) CCDF of follows (b) CCDF of shared media

Table 3. Comparison between Monitored users of BullyAlert and Instagram population collected in [18]

In continuing the narrative, we also put forth a detailed analysis of activities of other people in the user's profile, for example, the likes and comments received in the shared media sessions. Table 4a and 4b show the CCDF of the number of likes and comments received for the media sessions for both set of Instagram users. It can be seen that the media sessions shared by the users being monitored



(a) CCDF of likes (b) CCDF of comments

Table 4. Comparison between Monitored users of BullyAlert and Instagram population collected in [18]

are far less active in terms of getting likes and comments than their general counterparts. *This further solidifies our system perspective that our application's classifier will have fewer data to take care for, thus the classifier does not have to be as lightweight as described in [25], based on our current crop of data.* We acknowledge that the current crop of data is not enough to make a decision, so we plan to keep collecting the data to solidify this preliminary insight.

5 Conclusion

In this paper, we make the following contributions. First, we outline the motivation and design of a mobile application, BullyAlert, that adapts itself according to individual tolerance level for cyberbullying of the guardian. Second, we present a thorough architecture description of the components implemented to develop BullyAlert. Third, we provide a preliminary user analysis of both the guardians and the users being monitored by the application, and in the process, present several potential system issues/ challenges/perspectives using our current crop of data. In the future, we plan to expand our study, collection, and analysis of the guardian data.

References

1. `https://github.com/RahatIbnRafiq/AndroidCodesForCyberbullying`, [Online; accessed October 22, 2018.]
2. Broderick, R.: 9 teenage suicides in the last year were linked to cyberbullying on social network ask.fm. `http://www.buzzfeed.com/ryanhatesthis/a-ninth-teenager-since-last-september-has- committed-suicide` (2013), [Online;accessed 14-January-2014]

3. Center, C.R.: Cyberbullying research center. http://cyberbullying.us (2013), [Online; accessed September, 2013]
4. Council, N.C.P.: Teens and cyberbullying (2007), executive summary of a report on research conducted for National Crime Prevention Council
5. Dadvar, M., Trieschnigg, D., Ordelman, R., de Jong, F.: Improving cyberbullying detection with user context. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) Advances in Information Retrieval. pp. 693–696. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
6. Dinakar, K., Jones, B., Havasi, C., Lieberman, H., Picard, R.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying (Sep 2012). https://doi.org/10.1145/2362394.2362400, http://doi.acm.org/10.1145/2362394.2362400
7. DiProperzio, L.: Cyberbullying applications. http://www.parents.com/kids/safety/internet/best-apps-prevent-cyberbullying/ (2015), [Online; accessed February 6, 2015.]
8. E. Menesini, A.N.: Cyberbullying definition and measurement. some critical considerations. Journal of Psychology **217**(4), 320–323 (2009)
9. Goldman, R.: Teens indicted after allegedly taunting girl who hanged herself, bbc news. http://abcnews.go.com/Technology/TheLaw/teens-charged-bullying-mass-girl-kill/story?id=10231357 (2010), [Online; accessed 14-January-2014]
10. Hinduja, S., Patchin, J.W.: Cyberbullying research summary, cyberbullying and suicide (2010)
11. Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q., Mishra, S.: Towards understanding cyberbullying behavior in a semi-anonymous social network. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 244 – 252. IEEE, Beijing,China (2014)
12. Hosseinmardi, H., Rafiq, R.I., Han, R., Lv, Q., Mishra, S.: Prediction of cyberbullying incidents in a media-based social network. In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, San Francisco,CA,USA (2016)
13. Hunter, S.C., Boyle, J.M., Warden, D.: Perceptions and correlates of peer-victimization and bullying. British Journal of Educational Psychology **77**(4), 797–810 (2007)
14. Kontostathis, A., West, W., Garron, A., Reynolds, K., , Edwards, L.: Identify predators using chatcoder 2.0. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
15. Kontostathis, A.: Chatcoder: Toward the tracking and categorization of internet predators. In: PROC. TEXT MINING WORKSHOP 2009 HELD IN CONJUNCTION WITH THE NINTH SIAM INTERNATIONAL CONFERENCE ON DATA MINING (SDM 2009). SPARKS, NV. MAY 2009. (2009)
16. Kowalski, R.M., Limber, S., Limber, S.P., Agatston, P.W.: Cyberbullying: Bullying in the digital age. John Wiley & Sons, Reading, MA. (2012)
17. Lepage, E.: Instagram statistics. http://blog.hootsuite.com/instagram-statistics-for-business/ (2015), [Online; accessed February 6, 2015.]
18. Li, H.H.S., Yang, Z., Lv, Q., Han, R.I.R.R., Mishra, S.: A comparison of common users across instagram and ask.fm to better understand cyberbullying. In: 2014 IEEE Fourth International Conference on Big Data and Cloud Computing. pp. 355–362. IEEE, Sydney, Australia (Dec 2014). https://doi.org/10.1109/BDCloud.2014.87

19. Mohsin, M.: 10 youtube stats every marketer should know in 2019. https://www.oberlo.com/blog/youtube-statistics (2019), [Online; accessed Sep 6, 2019.]
20. Nahar, V., Unankard, S., Li, X., Pang, C.: Sentiment analysis for effective detection of cyber bullying. In: Web Technologies and Applications, pp. 767–774. Springer (2012)
21. Olweus, D.: Bullying at school: What we know and what we can do (1993)
22. Patchin, J.W., Hinduja, S.: An update and synthesis of the research. Cyberbullying Prevention and Response: Expert Perspectives p. 13 (2012)
23. Python, S.L.: http://scikit-learn.org/stable/modules/sgd.html, [Online; accessed October 22, 2018.]
24. Rafiq, R.I.: https://github.com/RahatIbnRafiq/cybersafetyapp_servercodes, [Online; accessed October 22, 2018.]
25. Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S.: Scalable and timely detection of cyberbullying in online social networks. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing. pp. 1738–1747. SAC '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3167132.3167317, http://doi.acm.org/10.1145/3167132.3167317
26. Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., Mattson, S.A.: Careful what you share in six seconds: detecting cyberbullying instances in vine. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. pp. 617–622. ACM, Paris,France (2015)
27. Rajadesingan, A., Zafarani, R., Liu, H.: Sarcasm detection on twitter: A behavioral modeling approach. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. pp. 97–106. WSDM '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2684822.2685316, http://doi.acm.org/10.1145/2684822.2685316
28. Sanchez, H., Kumar, S.: Twitter bullying detection. In: NSDI. pp. 15–15. USENIX Association, Berkeley, CA, USA (2012)
29. Smith, P.K., del Barrio, C., Tokunaga, R.: Principles of Cyberbullying Research. Definitions, measures and methodology, Chapter: Definitions of Bullying and Cyberbullying: How Useful Are the Terms? Routledge (2012)
30. Smith-Spark, L.: Hanna smith suicide fuels calls for action on ask.fm cyberbullying, cnn. http://www.cnn.com/2013/08/07/world/europe/uk-social-media-bullying/ (2013), [Online;accessed 14-January-2014]
31. Thom, B., Thom, B., Dhawan, A., Edwards, L., Dougherty, J.P., Coleman, R.: Safechat: Using open source software to protect minors from internet
32. Ueland, S.: 10 new social networks for 2019. https://www.practicalecommerce.com/10-new-social-networks-for-2019 (2019), [Online; accessed Sep 6, 2019.]
33. Villatoro-Tello, E., Jurez-Gonzlez, A., Escalante, H.J., y Gmez, M.M., Pineda, L.V.: A two-step approach for effective detection of misbehaving users in chats. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF (Online Working Notes/Labs/Workshop). CEUR Workshop Proceedings, vol. 1178. CEUR-WS.org (2012)
34. Willard, N.: Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress. Champaign, IL: Research (2007)
35. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB **2**, 1–7 (2009)