ISSN 1870-4069

Cyberbullying Detection in Social Networks: A Multi-Stage Approach

Laura P. Del Bosque, Sara E. Garza

Universidad Autónoma de Nuevo León
Facultad de Ingeniería Mecánica y Eléctrica,
San Nicolás de los Garza, NL, Mexico

{laura.delbosquevg,sara.garzavl}@uanl.edu.mx

Abstract. Cyberbullying, defined as the violent harassment of an individual towards a victim in electronic media, is a serious problem nowadays. If not prevented or mitigated, it can lead to affective disorders, poor academic performance, problems in social relationships, and — ultimately — to suicide attempts in youngsters and children. Because manual supervision in social networks (a space where cyberbullying can naturally occur) is laborious, automated approaches for cyberbullying detection are desirable. However, a considerable number of approaches treat this problem as merely aggressive text message detection, without considering the frequency of the harassment. The approach proposed in this work, in contrast, views cyberbullying detection as a process of three consecutive stages: aggressive text message detection, alleged aggressor and victim detection, and cyberbullying case detection. This approach was tested using a dataset extracted from Twitter; the approach obtained an F-score of 0.947 for the positive (cyberbullying) case.

Keywords. Cyberbullying detection, data mining, text mining, machine learning, sentiment analysis.

1 Introduction

Information technologies allow users to communicate with each other efficiently and to share their ideas and resources mutually. This, at the same time, shortens physical distance and enables collaboration, among other benefits. However, technology also presents a negative face by permitting misbehavior — and even cruelty — to occur more often. A vivid example of the former is given by cases of *cyberbullying*.

Bullying, which was traditionally limited to face-to-face encounters among children and youngsters in school yards, has unfortunately made its entrance into social media under the modality of cyberbullying [37]. Cyberbullying is a malicious, deliberate, repetitive act caused with electronic text messages [22], the main differences with respect to traditional bullying being the use of technology, the capability of reaching a great audience, the lack of time limits in the aggression, and the possibility of the aggressor to feel protected behind the technology. All of this makes cyberbullying more aggressive than traditional bullying [35], and — primarily due to the unsupervised use of technology — gives it the potential to easily get out of control [2], specially in social networks [30], whose rapid growth encourages this pernicious phenomenon [19].

Cyberbullying has been associated with negative experiences, as different studies have shown that victims report a poor academic performance, problems with family and other social relationships, and affective disorders [17, 36, 18]. In the worst case, cyberbullying can lead to suicide attempts when victims cannot cope with emotional distress due to the experienced abuse, aggression, and humiliation [3, 5]. In that sense, cyberbullying can lead to serious consequences that should not be underestimated.

Considering that cyberbullying is *an aggressive intentional act performed by an individual or group of individuals using electronic media to repeatedly contact a defenseless victim* [14], there are three main components in this phenomenon: an aggressive act (object), actors (aggressor-victim), and repetition (a pattern). Even though these three components seem to be present in several definitions [31, 22, 14], works tend to treat cyberbullying detection as a mere problem of aggressive text detection [1, 28, 7, 10, 11] by jumping to the conclusion that, if an aggressive act is taking place, then a harassing pattern has developed. The proposed approach, instead, views cyberbullying detection as a multi-stage *process* and attempts to identify this phenomenon in social networks using several progressive steps: (1) aggressive text message identification, (2) candidate bully and victim identification, and (3) cyberbullying case identification. The first stage relies on a profanity-based unsupervised technique, while the second one uses aspects from graph theory, and the third one is based on outlier detection.

The remainder of this work is organized as follows: Section 2 introduces relevant notions in network theory and Twitter, which is the case study for this work; Section 3 presents related work, Section 4 describes the proposed approach, Section 5 presents experiments and results, and Section 6 discusses conclusions and future work.

2 Background

The current section aims at briefly describing necessary concepts and notation from graph theory and Twitter. While graph theory is essential for social network analysis, understanding basic concepts from Twitter will serve for understanding the case study of this work.

2.1 Graph Theory

Networks are mathematically represented with *graphs*. A graph $G = (N, E)$ is a collection of entities (called *nodes*) and connections (called *edges*). Whenever edges are not bidirectional — that is, the presence of an edge (u, v) does not guarantee the presence of the edge (v, u) — the graph is said to be *directed*. If the edges contain numerical labels, the graph is said to be *weighted* as well. The *degree* of a node is the number of edges attached to this node; in a directed graph, the *in-degree* is the number of edges that enter a node and the *out-degree* is the number of edges that leave a node. A *multigraph* is a graph where multiple edges involving the same pair of nodes are possible.

A *subgraph* $G_s = (N_s, E_s)$ is a portion of a graph where $N_s \subset N$ and $E_s \subset E$. An *edge-induced* subgraph contains all edges in E_s and the nodes attached to these edges.

2.2 Twitter

Twitter[1] is a popular online social networking site where users are available to post short messages termed *tweets*. A user can *follow* other users, this meaning that the user will see the tweets from the followed users in his (her) personal webpage (called *timeline*). Tweets can be *directed* towards a specific set of users by including the addressed user names in the tweet; in Twitter, user names are preceded by the "@" (for example `@ausername`).

3 Related Work

Since the proposed approach covers three successive stages (isolated messages, actors, and cyberbullying cases through message histories), the discussed related work covers either of these stages and is either devoted to cyberbullying or to a similar phenomenon (e.g. pedophilia). For a more comprehensive review on approaches for cyberbullying automated detection, the reader is referred to the work by Salawu et al. [29] (in an *online-first* format to this date).

As mentioned in Section 1, cyberbullying detection has been recast as an aggressive text detection problem. This problem, in turn, has been mostly treated as a profanity detection problem (which is actually a simplification), where profanity is understood as the utilization of foul language. In both scenarios, the detection has been leveraged by machine learning and sentiment analysis techniques. One of the first approaches to address the aggressive text detection problem — outside the context of cyberbullying — was the Smokey system [34], a rule-based system used to identify hostile messages (also called "flames") in e-mail. Other outstanding approaches include the use of classifier ensembles [4], multi-step classifiers [27, 33], and the use of different sets of features — such as statistical, semantic, and linguistic [15]. Within the context of cyberbullying, Al-Garadi et al. [1], Ptaszynski et al. [26], and Yin et al. [38] propose basic machine learning techniques with different features, such as content and sentiment; similarly, Reynolds et al. [28] attempt to extract features to detect cyberbullying. Dinakar et al. [11], in contrast, train different topic-specific classifiers (race, sexuality, religion, intelligence, and physical attributes) to detect the aggression.

[1] Available at `http:\\twitter.com`.

In a later work, Dinakar et al. [10] propose the use of common sense reasoning to treat the case of aggression without the use of profanity.

There are also works that take into account the set of actors involved. For example, Chisholm [6] reports that women with aggressive communication styles tend to exclude their target victim and start conspiring against this victim, while aggressive men tend to use threatening words and phrases more frequently towards their victims. Furthermore, Dadvar et al. [7, 8] report that men and women use foul words with different frequency (men use certain words and women use other words); based on this finding, the authors implement different classifiers based on gender to detect aggressive text. Nahar et al. [19] present a two-step approach for cyberbullying detection, which starts with aggressive message detection and then utilizes the network-based HITS algorithm [16] to detect the actors involved.

Other types of works examine conversations or message histories to detect different types of phenomena. For example, Potha et al. [25] focus on sexual cyberbullying with a methodology based on time-series (a previous work on this approach being given by Potha and Maragoudakis [24]), where histories are represented symbolically and aligned with the aim of detecting a predator pattern. Also, Peersman et al. [23] use a three-stage system to detect cases of pedophilia; other works attempt to characterize pedophile conversations by means of features [12, 21].

4 Approach

The proposed approach aims at identifying cases of cyberbullying in social media by employing a set of successive stages. These stages arise from the definition of *bullying* given by Smith et al. [31]: *An aggressive intentional act performed by an individual or group of individuals repeatedly against a victim.* From this definition, three main components can be highlighted: aggressiveness, actors involved (aggressor-victim), and the repetition of the aggression. In that sense, cyberbullying can be seen as composed by *objects* (aggressive messages), *subjects* (aggressors and victims), and a *pattern*. Taking all of the former into account, the proposed approach consists of three stages:

1. Aggressive message detection
2. Alleged aggressor and victim detection
3. Cyberbullying case detection

The first stage concerns the identification of messages with aggressive content (media other than text remaining as future work) in a *social network*. A social network is considered as a structure composed by individuals that share a cybernetic relationship and have the capability of sending personal messages to each other [19]. At this stage, who sends and who receives the message and with what frequency is irrelevant.

The second stage, which is fed from the results of the first one, concerns the identification of alleged aggressors and victims. The former are defined as sending aggressive messages and the latter, conversely, as receiving aggressive messages. The message histories of these subjects are then more thoroughly analyzed.

The third stage, which takes in the subjects identified at the second stage, concerns the identification of cyberbullying cases by analyzing the frequency and intensity at which alleged aggressors harass alleged victims. This intensity is compared against the intensity of other conversations to search for an abnormal or outlier pattern. If a cyberbullying case is detected, the aggressor, victim, messages, and dates of these messages are returned as output.

The proposed approach uses a sampling technique similar to snowball sampling [20], since a set of messages is extracted from the social network according to several criteria (explained in Section 5); from these messages, a number is selected by the detection algorithm and sampled further according to other criteria (belonging to a certain user or pair of users). This sampling technique, along with the three stages, is explained in the following.

4.1 Aggressive Text Message Detection

The detection of aggressive text messages was performed using the approach proposed by Del Bosque and Garza [9], which concerns an unsupervised lexicon-based, term-counting strategy that identifies profane words. In summary, this approach, given a set M of messages, assigns a score sc_i to each message m_i. Such approach, even though simple, was selected to work around the *imbalanced class* problem; moreover, the aggressiveness score is be necessary for the second and third stages. This approach, in addition, has shown to yield satisfactory results for cyberbullying detection.

4.2 Alleged Aggressor and Victim Detection

In the second stage of the approach, the message sender and receiver become relevant. In that sense, a user that sends messages with a particular frequency and aggressiveness score is considered as an alleged aggressor or bully, while a user that receives messages with a particular frequency and aggressiveness score is considered as an alleged victim. For this case, that particular frequency is two or more messages within M (considering repetition) and that particular aggressiveness score is $sc_i \geq 5$ (considering this is the middle point of the scale used by Del Bosque and Garza [9]).

Formally, if the social network is treated as a directed multigraph (see Figure 1), $E = M$ and N is the subgraph of users induced by E — in other words, each node of the graph is a user of the social network and the users are connected by the messages they direct to each other (only the sample gathered is visible). The weight ω_i of each edge corresponds to its aggressiveness score such that $\omega_i = sc_i$. Let $\mathrm{deg}_{\mathrm{in}}(v, \alpha)$ denote the number of incoming edges to node v where $sc_i > \alpha$, i.e. $\mathrm{deg}_{\mathrm{in}}(v, \alpha) = |\{e_i : sc_i > \alpha, e_i = (u, v)\}|$ and, conversely, let $\mathrm{deg}_{\mathrm{out}}(v, \alpha)$ denote the number of outgoing edges from v where $sc_i > \alpha$, i.e. $\mathrm{deg}_{\mathrm{out}}(v, \alpha) = |\{e_i : sc_i > \alpha, e_i = (v, u)\}|$. Consequently, the set A of alleged aggressors is defined as $A = \{v : \mathrm{deg}_{\mathrm{out}}(v, \alpha) = \beta\}$ and the set V of alleged victims is defined as $V = \{v : \deg_{\text{in}}(v, \alpha) = \beta\}$, where $\alpha = 5$ and $\beta = 2$ as previously stated. Aggressor-victim relationships are then formed by extracting the possible aggressor for each $v \in V$ and the possible victim for each $a \in A$ such that $R = \{(a, v) : a \in A \vee v \in V\}$. Note that, while several of these possible aggressors and victims may not be in the A or V sets, in the third stage a deeper analysis will confirm or deny the existence of a cyberbullying case.

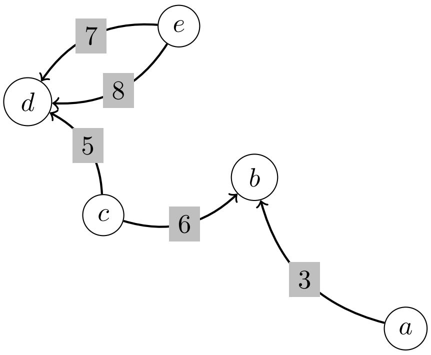

Fig. 1. Directed multigraph where nodes represent users and edges represent message aggressiveness scores. In this case, d is an alleged victim and $\{c, e\}$ comprises a set of alleged aggressors.

4.3 Cyberbullying Case Detection

While the first stage of the proposed approach works at the message level and the second one works at the user level, the third stage works at the historic level. A *message history* H_{au}, in this case, is defined as the set of messages directed from an alleged aggressor a to a user u in date d: $H_{au} = \{m_i : m_i = (a, u)_i, (a, u)_i \in E\}$ (date can be expressed as a function $f_d(m_i) = d$). In the third stage, an aggressive pattern (abnormal situation that spans for a time period) is searched for between an alleged aggressor and an alleged victim. When this pattern exists, a cyberbullying case is confirmed.

To detect an aggressive pattern, first the message histories between the alleged aggressor and each of the aggressor's contacts (including the alleged victim) are fetched for a period of preceding k time units (the selected k being six months, considering this is a reasonable amount of time for harassment to take place and also due to social network API restrictions). Next, an aggressiveness average score is calculated per history:

$$\bar{sc}(a, u) = \frac{\sum_{i=1}^{i=|H|} sc_i}{|H|}, \tag{1}$$

where $H = H_{au}$. Let also $\bar{sc}(a, v)$ correspond to a score where v is a possible victim.

Considering that an aggressor's contacts conform a set C_a, the average score for an alleged aggressor is given by:

$$\bar{sc}(a) = \frac{\sum\limits_{u \in C_a} \bar{sc}(a, u)}{|C_a|}. \tag{2}$$

Then, if $\bar{sc}(a, v) > \bar{sc}(a)$, a case of cyberbullying is detected between a and v where a is the aggressor, v is the victim, and H_{au} contains the messages and dates of the case.

For example, assume that $C_a = \{b, c, v\}$, $H_{av} = \{m_1, m_2, m_3\}$, $sc_1 = 6$, $sc_2 = 8$, and $sc_3 = 9$; then, $\bar{sc}(a, v) = \frac{6+8+9}{3} = 7.66$. Let $\bar{sc}(a, b) = 3.5$ and $\bar{sc}(a, c) = 4$. Then $\bar{sc}(a) = \frac{7.66+3.5+4}{3} = 5.05$. Since $7.66 > 5.05$, a cyberbullying case is reported between a and v with history H_{av}.

5 Experiments and Results

The purpose of the experiments was to detect cyberbullying instances by means of the multi-stage approach and validate these instances using evaluators. Experiments were performed using Twitter, which is a popular social network prone to cyberbullying attacks due to its public, unsupervised nature.

5.1 Setup

A message dataset was collected from Twitter using the API and the methodology described in the works by Del Bosque and Garza [9] and Escalante et al. [13], which in summary consists of obtaining directed messages with seed words (aggressive words, usually) and manually annotating these messages. An extensive list of aggressive words (shown in Table 1) was used to collect messages; these words were gathered with the aid of a native English speaker. A total of 13,313 messages was collected for this dataset. With the proposed approach, it was possible to detect ten alleged aggressors, whose message histories were tracked for six months. With these histories, cyberbullying cases were detected using Eqs. 1 and 2 (an example of a case is presented in Table 2).

Table 1. Aggressive words used to gather messages for cyberbullying cases.

c*nt	wh*re
punk as* b*tch	f*ggot
f*cking f*ggot	f*cking sl*t
f*cking c*nt	motherf*cker

Table 2. Cyberbullying case (bully: @user1, victim: @user2).

Message	Date
@user2 fat as*	May 14
@user2 f*cking fa*g*t sh*t bruh	May 14
@user2 sup wetback	May 1st.
@user2 DUMB*SS WETBACK IM CALLING THE FBI	April 16

A group of these detected cases was presented to a set of nine evaluators (according to Snow et al. [32], a minimum of seven non-expert evaluations are required to emulate the evaluation by an expert); collections of messages not detected by the approach as cyberbullying were also handled to the evaluators, hence the group of cases contained both *positive* (cyberbullying) and *negative* (no cyberbullying) instances. Note that manual annotation is being made *after* the instances are being detected and not *before*, as it is usual in machine learning training data. Therefore, we are validating that what the approach said was positive (case of cyberbullying) is actually marked as positive by a set of evaluators and what the approach did not detect as positive (no cyberbullying) is marked as negative by the evaluators. A total of 26 instances, where 18 were positive and 6 negative, was presented to the evaluators. To obtain the global class of an instance, the majority vote was taken into account; therefore, if an instance was voted as positive by eight evaluators, then the instance was globally classified as a cyberbullying case by the evaluators.

5.2 Results and Discussion

Table 3. Confusion matrix.

		Classified		
		as Positive	as Negative	Total
Class	Positive	18	2	20
	Negative	0	6	6
	Total	18	8	26

Table 3 presents the confusion matrix obtained from the evaluation. As it can be observed from this matrix, the results obtained by the approach closely match the manual evaluation; as a consequence, precision, recall, and F-score (F_1) present considerably high values.

Even though the results look promising, it is also important to keep in mind that this is only a small fraction of a vast social network that receives thousands of messages from thousands of users on a daily basis, thus making this kind of approach like looking for a needle on a haystack. It is possible to miss cases and, consequently, loose recall. However, on the other hand, being able to track

Table 4. Result evaluation.

Class	Precision	Recall	F_1
Positive	0.9	1	0.947
Negative	1	1	1
Average	0.95	1	0.97

these *real* cases on this haystack, regardless of the number of cases, should not be overlooked.

6 Conclusions and Future Work

A multi-stage approach for cyberbullying detection was presented. The first stage identifies aggressive text messages with an unsupervised, profanity-based algorithm; the second stage uses results from the first stage, as well as concepts from graph theory to identify alleged aggressors and victims, and the third stages analyzes alleged aggressor message histories and uses outlier detection to identify cyberbullying cases. The aggressive text message detection algorithm and the multi-stage approach were both validated through a set of experiments; while the algorithm is competitive against different techniques, the multi-stage algorithm is promising for further development. With regard to this, several lines of future work are possible. On one hand, it would be desirable to test the approach against similar methods to have a wider view on its effectiveness. Also, supervised methods could be tested for the aggressive message detection stage (a challenge here would be to address the class imbalance problem), since these methods showed good results. Moreover, techniques such as time series, Big Data, and deep learning could be incorporated into the approach to make it more scalable, robust, and capable of handling large amounts of data. Finally, a series of applications could take advantage from either the overall approach or individual stages; for example, an add-on in social networks could be inserted to warn users before posting an aggressive comment (i.e. the application would ask users if they are sure they want to post the comment, given that a high level of aggressiveness was detected on this comment).

References

1. Al-Garadi, M.A., Varathan, K.D., Ravana, S.D.: Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior 63, 433–443 (2016)
2. Alvarez-García, D., Nuñez Perez, J.C., Dobarro González, A., Rodríguez Pérez, C.: Risk factors associated with cybervictimization in adolescence. International Journal of Clinical and Health Psychology 15(3), 226–235 (SEP 2015)
3. Bauman, S., Toomey, R.B., Walker, J.L.: Associations among bullying, cyberbullying, and suicide in high school students. Journal of Adolescence 36(2), 341–350 (2013)

4. Bchir, O., Ismail, B., Maher, M.: Verbal offense detection in social network comments using novel fusion approach. AI Communications 28(4), 765–780 (2015)
5. Campbell, M.A.: Cyber bullying: An old problem in a new guise? Australian Journal of Guidance and Counselling 15(1), 68–76 (2005)
6. Chisholm, J.F.: Cyberspace violence against girls and adolescent females. Annals of the New York Academy of Sciences 1087(1), 74–89 (2006)
7. Dadvar, M., de Jong, F., Ordelman, R., Trieschnigg, R.: Improved cyberbullying detection using gender information. In: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012). pp. 23–26. University of Ghent, Ghent (2012)
8. Dadvar, M., Trieschnigg, D., Ordelman, R., de Jong, F.: Improving cyberbullying detection with user context. In: Serdyukov, P., Braslavski, P., Kuznetsov, S., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 7814, pp. 693–696. Springer Berlin Heidelberg (2013)
9. Del Bosque, L.P., Garza, S.E.: Aggressive text detection for cyberbullying. In: Gelbukh, A., Espinoza, F.C., Galicia-Haro, S.N. (eds.) Human-Inspired Computing and Its Applications. pp. 221–232. Springer International Publishing, Cham (2014)
10. Dinakar, K., Jones, B., Havasi, C., Lieberman, H., Picard, R.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS) 2(3), 18–48 (2012)
11. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. The Social Mobile Web 11(02), 11–17 (2011)
12. Eriksson, G., Karlgren, J.: Features for modelling characteristics of conversations: Notebook for PAN at CLEF 2012. In: CLEF 2012 Evaluation Labs and Workshop, Rome, Italy, 17-20 September 2012 (2012)
13. Escalante, H.J., Villatoro-Tello, E., Garza, S.E., López-Monroy, A.P., y Gómez, M.M., nor Pineda, L.V.: Early detection of deception and aggressiveness using profile-based representations. Expert Systems with Applications 89(Supplement C), 99 – 111 (2017), `http://www.sciencedirect.com/science/article/pii/S0957417417305171`
14. Espelage, D.L., Swearer Napolitano, S.M.: Research on school bullying and victimization: What have we learned and where do we go from here?[mini-series]. School Psychology Review 32(3), 365–383 (2003)
15. Justo, R., Corcoran, T., Lukin, S.M., Walker, M., Torres, M.I.: Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. Knowledge-Based Systems 69, 124–133 (2014)
16. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM 46(5), 604–632 (Sep 1999)
17. Kowalski, R.M., Limber, S.P.: Psychological, physical, and academic correlates of cyberbullying and traditional bullying. Journal of Adolescent Health 53(1, Supplement), S13–S20 (2013), the Relationship Between Youth Involvement in Bullying and Suicide
18. Machmutow, K., Perren, S., Sticca, F., Alsaker, F.D.: Peer victimisation and depressive symptoms: can specific coping strategies buffer the negative impact of cybervictimisation? Emotional and Behavioural Difficulties 17(3-4), 403–420 (2012)
19. Nahar, V., Xue, L., Chaoyi, P.: An effective approach for cyberbullying detection. Communications in Information Science and Management Engineering 3(5), 238–247 (2012)

20. Newman, M.E.: Networks: An Introduction. Oxford University Press, Oxford (2010)
21. Parapar, J., Losada, D.E., Barreiro, A.: A learning-based approach for the identification of sexual predators in chat logs. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
22. Patchin, J.W., Hinduja, S.: Bullies move beyond the schoolyard: A preliminary look at cyberbullying. Youth violence and juvenile justice 4(2), 148–169 (2006)
23. Peersman, C., Vaassen, F., Van Asch, V., Daelemans, W.: Conversation level constraints on pedophile detection in chat rooms. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
24. Potha, N., Maragoudakis, M.: Cyberbullying detection using time series modeling. In: 2014 IEEE International Conference on Data Mining Workshop. pp. 373–382 (Dec 2014)
25. Potha, N., Maragoudakis, M., Lyras, D.: A biology-inspired, data mining framework for extracting patterns in sexual cyberbullying data. Knowledge-Based Systems 96, 134–155 (MAR 15 2016)
26. Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., Araki, K.: Machine learning and affect analysis against cyber-bullying. In: Proceedings of the 36th AISB. pp. 7–16 (2010)
27. Razavi, A., Inkpen, D., Uritsky, S., Matwin, S.: Offensive language detection using multi-level classification. In: Farzindar, A., Keselj, V. (eds.) Advances in Artificial Intelligence, Lecture Notes in Computer Science, vol. 6085, pp. 16–27. Springer Berlin Heidelberg (2010)
28. Reynolds, K., Kontostathis, A., Edwards, L.: Using machine learning to detect cyberbullying. In: 2011 10th International Conference on Machine Learning and Applications and Workshops. vol. 2, pp. 241–244 (Dec 2011)
29. Salawu, S., He, Y., Lumsden, J.: Approaches to automated detection of cyberbullying: A survey. IEEE Transactions on Affective Computing Pending(Pending), 1 (2018)
30. Slonje, R., Smith, P.K., Frisén, A.: The nature of cyberbullying, and strategies for prevention. Computers in Human Behavior 29(1), 26–32 (2013), including Special Section Youth, Internet, and Wellbeing
31. Smith, P.K., Slee, P., Morita, Y., Catalano, R., Junger-Tas, J., Olweus, D.: The nature of school bullying: A cross-national perspective. Psychology Press, London (1999)
32. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In: Proceedings of the conference on empirical methods in natural language processing. pp. 254–263. ACM, Nueva York, United States (2008)
33. Sood, S.O., Churchill, E.F., Antin, J.: Automatic identification of personal insults on social news sites. Journal of the American Society for Information Science and Technology 63(2), 270–285 (2012)
34. Spertus, E.: Smokey: Automatic recognition of hostile messages. In: AAAI/IAAI. pp. 1058–1065 (1997)
35. Sticca, F., Perren, S.: Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. Journal of youth and adolescence 42(5), 1–12 (2012)
36. Tokunaga, R.S.: Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in Human Behavior 26(3), 277–287 (2010)

37. Walrave, M., Heirman, W.: Cyberbullying: Predicting victimisation and perpetration. Children & Society 25(1), 59–72 (2011)
38. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on web 2.0. In: Proceedings of the Content Analysis in the WEB. pp. 1–7 (2009)