Using Computer Technology to Address the Problem of Cyberbullying

Cheriton School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada
N2L 3G1
rcohen@uwaterloo.ca

ABSTRACT
The issue of cyberbullying is a social concern that has arisen due to the prevalent use of computer technology today. In this paper, we present a multi-faceted solution to mitigate the effects of cyberbullying, one that uses computer technology in order to combat the problem. We propose to provide assistance for various groups affected by cyberbullying (the bullied and the bully, both). Our solution was developed through a series of group projects and includes i) technology to detect the occurrence of cyberbullying ii) technology to enable reporting of cyberbullying iii) proposals to integrate third-party assistance when cyberbullying is detected iv) facilities for those with authority to manage online social networks or to take actions against detected bullies. In all, we demonstrate how this important social problem which arises due to computer technology can also leverage computer technology in order to take steps to better cope with the undesirable effects that have arisen.

Categories and Subject Descriptors
K.3.1 [Computers and Education]: [Computer Uses in Education]; K.4.2 [Computers and Society]: [Social Issues]

General Terms
Design, Management

Keywords
Cyberbullying, Education, Social Network, Online Safety

1. INTRODUCTION
In this paper, we discuss the social problem of cyberbullying and then present a multi-faceted approach to try to mitigate its effects, one that is grounded in computer technology. This work arose from a set of group projects conducted as part of a computer science graduate course at the University of Waterloo on Technological Solutions for Social Problems of Computers. We begin with an overview of the cyberbullying problem (including examples that have arisen in Canada). After this, we present some proposed technology to combat the problem, clarifying the parties that are engaged and the tools that can be offered. We conclude with a reflection on the possible value of the technology that we have designed, along with suggested steps for the future.

2. OVERVIEW
With the advent in technology and rise in social media, the lives of people around the globe have become much easier. Connectivity has reached a new level and this has allowed people to be in touch with anybody anytime. Along with the positive aspects of fostering relationships, this wave of unintercepted communication has also posed a serious threat to the personal lives and security of the people. Cyberbullying is a prime example. Dan Olweus, professor at psychology at Bergen University in Norway and a pioneer in research on bullying in his report of the Nova Scotia Task Force on Bullying and Cyber-bullying said that “The underbelly of the technology beast is dark and it has led to a growth in bullying. The immediacy and broad reach of modern electronic technology has made bullying easier, faster, more prevalent and crueler than ever before” [10].

2.1 Cyberbullying
Cyberbullying is the oppression of other people using Internet-based technology. Harassing people through internet websites like Facebook, Twitter, emails, text messages and thus breaching into the personal zone of the victim is what a cyberbully does. Such bullying can be done by sending hate mails, abusive text messages, posting on the space on social networking websites, hacking into the accounts, getting access to the personal information of the victim and thus blackmailing him over it [3]. There may be many more ways of carrying out this undignified process. Some of the ways are described in Figure 1.

2.2 Effects of cyberbullying
Cyberbullying can leave a lasting impact on the victims or even impair them mentally. For example, stalkers can take the electronic way of teasing their victims and thus go to any extent to vilify them. The Supreme Court of Canada said that cyberbullying leads to fear, anxiety and loss of self-esteem, and that it is particularly harmful because content is spread quickly and anonymously. The impact of cyberbullying is directly proportional to the high speed of the Internet
Cyberbullying is different from other physical kinds of bullying as it can reach a large audience at one go. A study shows that of those who are victimized by cyberbullies, 22.8% had been scared for their safety [19]. The psychological impact on the victims is sometimes so much that they take the extreme steps of ending their lives. Many instances in the recent past prove this fact. Amanda Todd, a teenager living in Port Coquitlam, B.C., killed herself after she posted a video on You Tube summarizing the impact of online bullying and physical assault she had over the years. Jamie Hubbley, a 15-year-old gay teen from Ottawa, Ont., killed himself after 15 years of suffering from depression over the bullying he experienced as early as in Grade 7 [13].

2.3 Social actions to curb cyberbullying

It is highly important to spread awareness of the signs exhibited by victims of cyberbullying and to provide them with adequate support and intervention [10]. Certain factors should be kept in mind while formulating laws about cyberbullying, for example what effect cyberbullying has on the minds of the victims and how they react to it. Also, it should be kept in mind that any kind of online harassment or incident should be reported to higher authorities of the organizations and be taken care of by them [3]. For instance, if cyberbullying happens in schools, school administrators, staff, students, and the families of the victims should be taken into account. Victims of cyberbullying are not very comfortable with sharing their trauma with others because:

1) it is a humiliating experience and victims don’t want to publicize it; 2) victims fear that their peers as well as parents would blame the victims themselves, as they are socially active; 3) victims do not want to portray themselves as weak and vulnerable in front of others; and 4) bullies make victims feel insecure by barging into their personal space and blackmail them to make it public. So, people fall for this trap in the pretext of feeling safe otherwise [17]. In order to tackle cyberbullying, the Canadian government has introduced legislation to create Canada’s first cyberbullying investigative unit. This legislation would allow victims to get security from the court if harassed online [4]. The province of Nova Scotia has followed suit just recently with its own legislation [12].

With cyberbullying leaving such huge impacts on the lives of people and destroying their security, Prof. Olweus offers a simple suggestion for reducing cyberbullying in schools. He says that it would be beneficial to investigate a few cases and make the outcomes public so that the students would further be terrified to conduct such hateful behavior because of the actions taken against the bullies [10].

3. COMPUTER TECHNOLOGY TO ADDRESS CYBERBULLYING

The previous section outlines what citizens and governments have envisaged for taking actions to mitigate this new social problem. In this section, we present distinct computer technology for addressing cyberbullying. We begin with some proposed approaches for detecting cyberbullying, then follow this with methods for supporting the reporting of cyberbullying. After this, we discuss strategies for discouraging cyberbullying, including suggested actions to be taken by those in authority.

3.1 Detection of cyberbullying

The first step that can be taken to use technology to address cyberbullying is to examine messages that are arriving, to try to detect instances of bullying. Below we sketch two proposed approaches focused on an examination of incoming messages. The first enables a labeling of each message as a likely instance of bullying (which may then be used by recipients to selectively ignore unwanted content). The second enacts a filtering process, to remove from sight those messages deemed to be problematic.

3.1.1 A proposal to label malicious messages

We outline here one strategy that may be used to detect cyberbullying, intended for the context of social networks. The core suggestion is to model the reputation of each message, via feedback from users and to attach a warning label to those messages which may be instances of bullying.

On some social networks, the concept of rating messages either positively or negatively shapes the entire layout of what is seen by end users [2] [1]. On both Reddit and Stackoverflow, individual users receive a reputation score. The reputation score of an author we devise is essentially a function of how much other users appreciate viewing the messages they post. Here, we mean for a message to include any sort of content shared by a user. However, our solution relies partly on the use of text parsing, so it is not as effective for messages without text. If user Bob has messages $m_1, ..., m_n$, then each message has an overall message score, which we can denote as $r_1, ..., r_n$. For each message in the social network, users are able to either vote positively or negatively on them. Hence, the message score of Bob’s message $i$, $r_i$, depends on the number of positive votes and negative votes. This means that $r_i = \frac{numpos_{i} + 1}{numneg_{i} + 1}$, where $numpos_{i}$ and $numneg_{i}$ are the number of positive and negative votes on the message, respectively. To calculate Bob’s reputation score, we can simply sum over all of his message scores and divide that number by his total number of messages, $n$, taking the average of his message scores. $rep(Bob) = \frac{1}{n} \sum_{i=1}^{n} r_i$. If $rep(Bob) < 1$, then we assume Bob is more likely to be malicious, in proportion to how far away Bob’s reputation lies from 1. If $rep(Bob) > 1$, then Bob is more likely not
malicious, in proportion to how far past Bob’s reputation lies from 1.

In its current state, Facebook typically has its users submit messages which can be seen by any number of other users in their local social network, usually by their Facebook friends. Messages can be “liked” by other users, providing Facebook with a metric to determine a ratio of views to likes on a certain message. We propose a system that involves including both a positive and negative review of a message by its audience on personal social networks. We use Facebook to illustrate an example. By including such a system, we can aggregate the positive and negative reviews of posts in order to help determine what the posting user’s reputation score should be. In order to facilitate detection of harassing content, machine learning can be incorporated to flag messages as being likely harassing [8], which can be used to further negatively impact a posting user’s reputation score. By incorporating a machine learning paradigm into the solution, we can begin to catch obvious maliciousness of users on the social network [8]. Including user feedback in the form of positive or negative reviews on messages helps to alleviate when the machine learning algorithm falls short. The work done by Dinakar et al. illustrates the effectiveness that a binary classifier can have in detecting obvious cyberbullying. Using user input helps to detect more complex harassment, such as sarcasm, which fails often in the model of [8]. One such example of complex harassment that could be caught by users could be a message such as: “You are a lot smarter than you look.” Some examples of what the message rating system could look like are below.

The message thread in Figure 2 displays the changes to Facebook’s current thread structure required to implement this system. By including two possible ratings instead of only “likes”, user input can be combined with the binary classification from [8] to determine the final outcome of the reputation score. In the above case, the machine learning algorithm determines that Bully1 has written a malicious message. However, for Bully2, it fails to determine that the message is malicious. As a second step to determining whether a message is malicious or not, user ratings are taken into account and thus determine Bully2 is being malicious since Bully2 has nearly twice the number of negative votes to positive votes. To reflect this, the same warning symbol is used on Bully2’s message. In this example, the votes on messages are hidden from users.

By cultivating a reputation score for users in the network, we can classify users who are more likely to engage in cyberbullying, as well as users who are more likely to become victims of bullying. Suppose for example a Facebook user has many friends with a negative reputation score. Since the reputation score provides a rough guideline as to whether a user is one who engages positively or negatively with their peers, we can see how having many friends with a negative reputation could increase the likelihood of being bullied. To assist such at-risk users, we can provide educational resources to help cope with bullying. More importantly may be to help these users expand their social network in a positive manner, by suggesting people they may know with a higher reputation score. Since our solution provides what we believe is a reflection of a user’s likelihood to be malicious, we can use this information to dampen the effects of conventional cyberbullying.

More specifically, data shared on the Internet is more easily duplicated and saved when compared to data in real life. This is important to note since victims of harassment on Facebook, for example, have an indefinite record of the messages they have received. This lies in contrast with the immediacy, as well as temporariness, of verbal harassment in a face-to-face confrontation. We see this as a point of concern because amassing an archive of such messages can be detrimental to the mental health of victims, such as Amanda Todd [6]. In order to address this, we look to our reputation score for insight.

Consulting with a user’s history on Facebook in our solution’s paradigm, we can subject messages that are likely malicious to an expiry date. Such an expiry date would only impact the visibility of the message, not necessarily destroy its storage, as it may be needed for future investigation. In doing so, we aim to dampen the effects of archived malicious messages.

It is important to note that with any kind of rating system, there is the concern of possibly deceptive ratings being provided, which would be exacerbated if, in addition, collusion existed. This would arise in our context, if a pool of peers sought to collectively protect a potential bully or discredit an unfortunate victim. A number of approaches
from the trust modeling literature in artificial intelligence could be useful to consider, for example the work of Kerr on collusion resistance [11]. Another valuable direction would be the promotion of methods inspired by community-based privacy protection, such as that of Garg et al. [9]. Some interesting suggestions from this work include convergence on community-based norms and the consequence of excluding those who do not adhere. The larger community thus provides an important level of protection. Leveraging a peer-produced privacy approach towards reputation modeling is also echoed in the work of Vicceco et al. [21] who state: “The same mechanisms that ensure that individuals retain control over their own information (privacy) make the reputation system more robust”. This is intricately tied to owner-enforced access control.

3.1.2 A proposal to filter suspected messages
Victims of cyberbullying can be harmed by the emotional effects of reading abusive messages directed at them. If the victim does not receive the abusive messages, they will not be harmed in this way. In this subsection, we propose to use a filtering mechanism to classify messages as “abusive” or “non-abusive” (or “positive” and “negative,” respectively). There have been a number of attempts to create an automatic detection mechanism for instances of cyberbullying [7, 14], but our proposed system appears to be unique in its use of a combination of customization for individual users and a trusted third party selected by each user.

Ideally, a filter could correctly classify all messages and discard insults and similar abusive messages. In a practical system, the filter will not be completely reliable; there will be false positives and false negatives in at least some cases. If messages labelled abusive were simply discarded, important non-abusive messages would likely be discarded as well. There is an additional problem in that some abusive messages may require measures other than simple deletion; specific threats, for example, may need to be reported to the police. It would be difficult to create an automated system to reliably recognize threats that should be reported to the police, and automatic reporting would be problematic (not least because it would be likely to waste police’s time).

The problem of false positives and the problem of discarding threats can both be dampened by diverting messages labelled abusive to a trusted third party. This third party is likely to be a parent or guardian for younger users; older users may choose a close friend instead. Since the messages are not directed at this third party, the third party would not be expected to suffer significant harm. If a message is correctly classified, the third party can simply delete the message or take further action if the contents of the message require it. In the case of a false positive, the third party can return the message to the filter, which would then deliver it to the intended recipient. Dadvar and de Jong suggested flagging messages in a public forum detected as possible cyberbullying incidents to a forum moderator [7], which is similar to but distinct from our proposed use of a user-selected third party in filtering private communications.

The filter can be designed to alter its classification function in response to feedback about false positives from the third party and false negatives (abusive messages that are mistak-

This filter is closely analogous to a spam filter. In this case, the objective is to exclude messages which are likely to cause emotional harm; in the case of a spam filter the objective is to exclude messages which attempt to sell the recipient something. Aside from the nature of the content that is to be excluded, the primary difference is that because reading the message alone is sufficient to harm the recipient when filtering abusive messages, so instead of a spam folder the flagged messages are sent to another user for review. Since the technical aspects are so closely related, the success of spam filters suggests that a similar filter for abusive messages would be effective.

The selection of features would be very important to a filter of this type. We propose three general types of feature that are likely to be useful: bag-of-words features, sender information, and sentiment analysis. The presence or absence of specific words in the message or the number of occurrences of each word (often referred to as a “bag-of-words” representation) is often used in spam filters, with considerable success because certain words are more common in spam than in legitimate messages, and vice-versa. Similarly, it is reasonable to believe that certain words would be indicative
of abusive and non-abusive messages. Sender information includes the name of the sender, and may also include information about the relationship between the sender and the user, if it is available; this type of information was used by Dadvar and de Jong [7]. This is roughly analogous to email header information, which has been found to be useful in spam filtering [23]. Sentiment analysis is more sophisticated than simply looking for specific words; it attempts to determine, at least in broad terms, the emotional content of a message [15]. It is a commonly-used technique in analyzing social media data, and has also been applied to automatically detect email “flames” [20], which is similar to detecting abusive messages. Sentiment analysis, though somewhat challenging, could prove very useful in combination with other methods. Previous work on detecting cyberbullying has made extensive use of explicit sentiment analysis [7, 8, 22].

Selection of a learning method is also vital to the success of a learning filter mechanism. One popular type of classifier in spam filtering is the naive Bayesian classifier [5, 23]. These classifiers are also popular in spam filtering, perhaps due to their combination of simplicity and sufficiently good performance [5]. For a filter for abusive messages where abusive and non-abusive messages may come from the same sender, it may be useful to relax the assumption of conditional independence given a true classification and the sender; this allows the system to learn the characteristics of abusive and non-abusive messages sent by individual users, and may increase the accuracy of classifications. A filter of this type would not be a traditional naive Bayesian classifier, but instead would be a (very simple) Bayesian network. Since the bag-of-words and sentiment analysis features would still be treated “naively,” the characteristics would probably be quite similar to naive Bayesian classifiers.

### 3.2 Reporting of Cyberbullying

Another critical step in the effort to combat cyberbullying is support for the reporting of instances of bullying. As will be discussed in this section, there are several avenues to pursue. We introduce a proposal to bring trusted third parties into the picture, as recipients of information which may then serve to shield the intended victim of the cyberbullying. We also outline a very detailed system which can be used by several parties (victims or onlookers or those who are looking after the victims) to record possible instances of malicious actions, which can then be investigated further by authorities. We finally discuss options to engage suspected bullies, in order to provide them with information that may serve to enlighten and to deter future misdeeds.

#### 3.2.1 A reporting system with third party assistance

We begin with some examples to motivate the design of our reporting system. Example 1: Bob is browsing his Facebook page and finds a post with his name mentioned. The post contains offensive words targeting at him. He believes he has been cyberbullied and decides to report. Currently, there’s no functionality on Facebook that enables such a report, which leaves him no option other than ignoring the post. Our solution proposes a Cyberbully Reporting page, where Bob can identify himself as the victim by pasting the link to the post where cyberbullying messages are found, and the name of the person where the post originates. This enables Facebook to detect cyberbullying messages and carry out corresponding actions. Example 2: Alice is browsing Facebook and finds a post that attacks her friend Bob with offensive words and graphics. She believes the post is to bully Bob and therefore decides to report it. Our system allows for someone other than the bully to register information.

We now describe our proposed reporting system. To address the issue of cyberbullying we offer explicit help support on Social Networking Systems (SNS) such as Facebook and Twitter. The idea is to gather cyberbullying intentions from user reports and to provide direct and immediate protection to the victims. The entrance of the helping process is a help icon which is prominently displayed on the top bar of the SNS websites such as Facebook. The user will get acknowledged that this help icon is aimed to help them to deal with the bullying problem online by a simple click. The figure below (Figure 4) explains the workflow of the help process. The whole workflow consists of 4 phases.

![Figure 4: Help process workflow diagram illustrating the details of the proposed reporting system.](image-url)
for future determination. Otherwise the system sends the victim a notification showing that he/she may possibly be suffering from the cyberbullying or harassment in the target post and asks whether the victim needs help. The victim confirmation process aims to reduce the false positives and affirm the report as an actual improper behavior which harms the victim. When the cyberbullying is confirmed, the process will split into two branches dealing with the victim and bullies respectively.

Victim Helping Phase. The system provides a list of features among which the victim can choose those best describing his/her current situation. The features would include: flaming, privacy leaking, online harassment, group bullying, and stalking. With the feature information the system decides the severity level of the current harassment and recommends solutions to the victim such as disable sharing of the post, message blacklist, and accessing legal aid.

Improper Online Behavior Notification Phase. The system also aims to take on the role of social media education by providing users with tips of proper online behavior. The system targets the badly behaved users with both victim reports and automatic detection. As shown in the workflow, after the cyberbullying confirmation the victim can report the bullies to the system. We note that it is challenging when a large group of people commit cyberbullying together. In this case, the system could possibly automatically analyze all the users’ comments for the target post and filter suspects. All suspected users reported by the victim and automatically detected by the system will get a notification from the system providing the target post URL and showing that their online behavior is considered improper and resulted in a bad effect on other users. The system keeps track of users’ bullying history. A threshold is set to alarm the users labelled as bullies who are reported frequently to mark them as dangerous users. Those users will get another notification with stronger words indicating that they are reported frequently and should care about their online behavior. The system administrator will manually look into the record of the dangerous suspects and decide whether the account should be constrained or banned.

Protection. In the stories described in [16] and [13], victims are repeatedly bullied over an extended period of time, which lead to the tragedies that could have been prevented if the bullies were stopped immediately once the incident started. Therefore the second portion of our proposed solution deals with post-bully protection of the victims. We focus on how to protect the victim from receiving further damages once the cyberbullying has been identified. Our proposed solution is to give the user the option to delete/block the post so that no one can edit or comment on the post. The means that fewer people will be able to see the post and the psychological impact to the user is minimized. We acknowledge the inherent challenge in deleting information that has previously existed online, the so-called “right to be forgotten” issue [18]. Notwithstanding this concern, there is in fact a growing trend to ensure that online providers admit options of erasing one’s past.

Monitoring. Another feature of our solution is the ability to warn cyberbullies about their actions. After the user identifies the bullying post, our system will send a message to the bully with a warning to notify and educate him. If the bully makes another post to attack the user again, he could receive a tougher penalty, including account disabling, and reporting to law enforcement units. This series of actions is aimed to monitor the behavior of the bullies to deter them from carrying out further bullying actions and hence to protect the victims.

3.2.2 A centralized platform for reporting
In this section, we outline an alternative reliable reporting platform to the victims or witnesses where they can report the incidents of cyberbullying. This platform is a web portal which could directly be under the control of an authoritative department such as police or any concerned government official. This initiative would require the authorities to work in close contact with the administrative teams of various websites. With the collaboration of the admin team and the police department, all the reported incidents can be tracked and hence the bullies can be accordingly punished. This would instill confidence in people that they have someone to look forward to in case they witness or experience any kind of cyberbullying. The Cyberbullying Reporting Platform (CBRP) is presented is this section.

User Side of CBRP. Figure 5 is a sample image of the system’s web portal. CBRP is a basic web portal where the victims or the witnesses can report the incident. They will be required to create a basic username on the website which would keep track of all the incidents related to that user. The form contains personal information like name, age and current university/college/school that the person is attending. The basic purpose of getting this information is to narrow down the scope of possibilities to find the bully. Further, the portal asks for information about the source of the incident like the website on which the unwanted content is published. The victim’s email ID and username is required and also that user ID (or preferably the email address) from which the messages are being sent. All this information would be made visible to the admin of the corresponding website to track down the provided IDs to verify if the bullying is being undertaken. In addition to these, there is a provision of uploading exact proofs in the form of images or messages being sent by the bully. All the images posted or the text messages delivered can be uploaded by the reporter and later used by the authorities to verify the bullying incident and to use as proof against the bully. Even the email headers (if possible) can also be used by the reporter, as the headers contain a lot of information about the mail sender. With an email header it becomes far easier to get one’s hands on the bully.

This form can be filled by the victims themselves or by any witness of cyberbullying. All the incidents will be under one thread with the created username. The Personal information will not be shared with the admin of websites. Only the information required for tracking and verifying will be shared. All the information will be reported to the higher
authority in charge, maybe the police department. In case the bully continues and the seriousness is grave, the police can use all the necessary information uploaded to track and catch the bully.

Acknowledgement Screen. After the incident is reported the user is presented with an acknowledgement screen. A snapshot of this screen is shown in Figure 6.

![Acknowledgement Screen](image)

Figure 6: Acknowledgement window after the cyberbullying incident is reported, shown to the victim or witness.

This screen solves various purposes like: 1) It provides encouragement to the person who has reported the incident. The immense courage of the person is appreciated; 2) It provides satisfaction to the person when he/she knows that administrators and the concerned police department will be looking into the case; 3) It provides hope that the case will be taken care of in a specified number of days. So this provides a big hope to the person who is being harassed; and 4) This page further motivates the person for providing any other information which could be helpful in catching the bully ASAP.

View for Authoritative Department. So far, we have seen the user’s side of the web page. Let us now have a look at the portal available to the police department or other authority in charge. A snapshot is shown in Figure 7.

![View for Authoritative Department](image)

Figure 7: CBRP view from the authority’s perspective; notice how the incidents are prioritized to help effective deal with cyberbullying.

The police department or any other authority in charge would have access to this page. They can get the list of incidents arranged by website names or by response date. The whole profile information along with all the materials uploaded under the same thread would be available to the department. In order to have a close monitoring on the response dates, they can directly contact the admin teams of the corresponding websites. This check would get things moving at the end of the websites’ admin team and will help to take action on time. It also provides all the proofs to the police so they can track the person in case further bullying continues from the same person. This could help in producing proofs in legal proceedings as well.

View for Website Admin Team. Similar to the police department section, there is an admin side of the portal as well. This would be available to the admin of the website to trace the incidents. A glimpse of the page is shown in Figure 8.

![View for Website Admin Team](image)

Figure 8: CBRP’s administrative portal. This enables administrators to manage the system so that it can be used by authority, witnesses, and victims.

The working of this whole concept drills down to the responsibility of the web admin of the website. All the necessary
information uploaded by the victim may be made available to the admin. The admin is responsible for first of all verifying the information uploaded by the victim. She has to make sure that the information provided is actually present on the website so that the incident reported is not a fake. Also, feedback can be sent to the victim at this point.

**Impact.** CBRP, if used actively, can minimize and can put a stop to the cyberbullying as soon as it starts on the web. Larger teams of professionals can be used to monitor the website and track down the bullies in no time. It will also be easier to put them under criminal trial with all the proofs collected against them. The more incidents this site would be able to stop, the more trust will be built and fewer incidents of cyberbullying will occur. In all, the courage of the people to report the incidents, together with the active participation of police and the admin teams of websites, can begin to address this important problem of cyberbullying and to see that bullies are brought to justice.

Note that our design for this tracking system will ultimately form a centralized database which will store all incident report information. This database will thus help in many aspects. For example, we can have a better view on the social network distribution of cyberbullying through profiling the URLs those reports provide. Different kinds of studies and analyses can also be conducted and processed by taking advantage of first-hand data in our database, provided as parts of the forms that are filled. Thus, the features and characteristics of cyberbullying on social networks can be further observed and studied.

### 3.2.3 Educating victims of cyberbullying

To help mitigate bullying, one option is to incorporate educational material to at-risk users. In this section, we sketch one such approach. We determine at-risk users by looking at the reputation of other users in their local social network, such as their friends. Presenting material that describes best practices for social network use can help to educate, including reminders to be mindful of the audience of a message, be aware of the content shared, and be cognisant of the perils of releasing personally identifiable information on the internet. This information could be displayed subtly, but more prominently to those who are most at-risk. Alongside these best practices, phone numbers for support hotlines in the area of the user could be incorporated to further support users left victimized. In Figure 9, we display such an example of educational material we can show to users.

### 3.2.4 Deterring cyberbullying

While there are various options for delivering education to victims, of course the bullies themselves merit some instruction as well. We describe an approach for deterring cyberbullying, actions which can be taken against bullies that will result in better educating them and in incentivizing them to desist from this kind of activity.

Although we want to determine when users are sharing messages that are malicious, we do not want to make prominent what is going on behind the scenes of our solution. In other words, publicizing the reputation score of users (as in Section 3.1.1), and alerting all viewers of a message of its maliciousness is probably not a good idea: this may further reinforce behaviour we are looking to discourage. To deter bullies from engaging in malicious behaviour, we can force users with low reputation to engage in educational tests and readings before accessing Facebook. Examples are presented in Figure 10 and Figure 11.

By creating overhead for those identified as bullies, it becomes in their best interest to either disengage in the network altogether, which is costly if most of their peers use the network, or to seek a higher reputation to avoid being forced to do such tasks. In other words, for a reputation score system, if we increase the ranking weight received from the machine learning algorithm, and decrease the weight given to user ratings, it becomes harder for user input to wholly shape reputation scores. When someone with a low reputation score logs in, they may be presented with a reading and a quiz to do, and perhaps score sufficiently well, before they can interact on the social network. Of course, additional sanctions can be put on users with a poor reputation, but these should be reviewed for efficacy before implemented.

### 3.3 Roles for authorities

From the proposals discussed to date there are a variety of roles to be played by parties other than the bullied, the bullies, and the onlookers. Those in trusted positions or in authority can be engaged to make a difference. Recipients of

---

**Dealing with Cyberbullying**

Cyberbullying typically has a detrimental effect on its victims. Victims often feel helpless and as a result suffer from depression, anxiety, and social isolation. There are many practices that you can take to prevent cyberbullying from happening to you or anyone else in your environment.

**How to react to cyberbullying**

- **Cut off the bully**—If the bully is making direct communication with you, tell them to stop. If he or she refuses to stop, block him or her from the communication channel he or she is using to harass you. Studies have shown that bullies typically bullying to seek attention and will often stop if they are ignored.

- **Record**—If the bully continues to harass you, keep records of all the communication, i.e. phone calls, messages, posts, emails, sent. If the bullying is physical as well, record the time of the event and what happened. For phone calls, dialing *123* before the end of a call will have the bully’s phone number recorded by the phone company. These records will serve as important evidence against the bully.

- **Reach out**—Report cyberbullying to someone in authority such as your administrators, teachers, or managers. You can also report cyberbullying to the police, as undesired repeated harassment is considered a criminal offense. It may also be helpful to talk to close friends and family for emotional support. There are also many helplines and counselors that you can reach out to for help.

- **Report to Service Provider**—Many service providers have terms of use agreements that its users are required to follow regarding decorum and etiquette. Reporting the cyberbullying incident can get them banned from the platform. Moreover, the service provider may also be able to track down the identity of anonymous bullies and remove defamatory content.

**What Not to Do**

- **Become a cyberbully yourself**—Sinking to the bully’s level will not help to solve the problem. You are only becoming a bully yourself and will make other suffer as you have.

- **Broadcast the message**—Do not forward or share the messages with others who are not aware of the situation. Messages forwarded to people who are not aware of context can exacerbate the problem greatly.

- **Let the bully get to you**—No one deserves to be bullied or harassed at all. The inappropriate behavior of bullies often has nothing to do with the victim. Bullies tend to be insecure people with problems who are taking it out on other people as a means of release. They are cowards who have no courage to deal with their own problems.

**References**


---

**Figure 9:** A sample document that may be shown to cyberbullying victims detailing support options available.
Cyberbullying

Cyberbullying is defined as the use of technology to support deliberate, hostile and hurtful behaviour towards an individual or group of individuals [2].

Why People Cyberbully

Just like other forms of bullying, cyberbullying is about gaining power and control. Those who bully others are trying to establish dominance over people they perceive to be weaker than them. While technology can be used as a positive communication tool it can also be used to hurt others [1].

In scientific studies, it has been found that people engage in cyberbullying activities to direct their frustration, anger, hurt, and difficulty they are experiencing elsewhere. Some also do so due to lack of attention from friends and family. Others bully to fit in with their friends, in cases of group bullying [1].

Impact of Cyberbullying

- Feel helpless, angry, depressed, and/or anxious
- Feel invade in cases that the bully is anonymous
- Feel shame and embarrassment in a worldwide venue
- Surprise at how communicate and content can be blown out of context
- Have a tendency to isolate oneself from social group
- Feel that harassment cannot be avoided because technology is easily accessible
- More susceptible to self-inflicted harm and even suicide

The Law

Some forms of online bullying are considered criminal acts. Under the Criminal Code of Canada it is a crime to:

- Communicate repeatedly with someone if the communication causes them to fear their own safety or the safety of others
- Write something that is designed to insult a person or likely to injure a person’s reputation by exposing them to hatred, contempt or ridicule.
- A person may also be violating the Canadian Human Rights Act, if he or she spreads hate or discrimination based on race, national or ethnic origin, colour, religion, age, sex, sexual orientation, marital status, family status or disability [3].

What to Do

You can be prosecuted for involvement in cyberbullying. Here are some tips to not be a cyberbully:

- Think before you click! Consider the recipient’s feelings before sending the message. Chances are if you would not say it to a person-to-person, then you should not be posting the message.
- If a group of your friends are cyberbullying an individual, do not participate. Notify an authority.
- Private messages between you and another person should not be publically shared.
- If you are bullying to seek attention or because of difficulties in your life, speak with an adult and seek the proper social support needed.

References


Figure 10: A sample document that may be shown to cyberbullies detailing potential consequences of their actions.

reports from social networking services can include both police and government authorities and technical administrators (in charge of the social network or website). A final valuable party to assist in coping with cyberbullying incidents is that of trusted third parties (e.g. parents or teachers) as outlined in Section 3.1.2. The proposal sketched there directs messages to these people as gatekeepers, so that the burden of the “negativity” is absorbed in this way. Of course, this is perhaps practical only when the message recipients are quite young and willing to forgo the privacy of their communication.

3.4 Technological solutions to cyberbullying

There are some of the ongoing technological challenges to introducing techniques from computer science in order to address cyberbullying.

With respect to possible message filters, two primary hurdles are preventing the removal of valuable messages when attempting to filter and getting around the loss of privacy, if third parties are introduced as part of the filtering solution. Suggested avenues for future refinement of these proposals include more clever semantic processing of messages for the first issue and perhaps tagging suspected messages as an intermediate step, for the second challenge (so that the recipient can approve the forwarding of specific messages only to others who may be assisting).

As for efforts on deterrence, one issue is the fact that those who are the perpetrators of bullying will eventually become jaded to any efforts to educate them or to enlighten them. One important reality is the fact that those who commit cyberbullying may already have a history of bullying outside of cyberspace [19]. All the same, we feel there will be some kind of residual effect if the bullies are consistently required to, for instance, demonstrate their competence in sensitivity to victims. For the future, an avenue that may show some promise is to vary the required task that cyberbullies must complete, with each new incident, to ensure that the engagement is continuously fresh and valuable.

Reporting of cyberbullying incidents encompasses a rich set of possibilities as outlined in this paper but even here there are opportunities to improve on what is being offered. Perhaps the greatest challenge is to ensure that incidents can be reported very quickly; we envisage an option similar to a Kids Help Line, as a button that can be clicked, clearly visible as an option on all social networks, for instance. First steps with this approach were outlined in Section 3.2.1.

This leads to perhaps the final general challenge with respect to cyberbullying and efforts to enable its reporting: the case of false reporting. While we have suggested various steps to try to cope with this, more effort will need to be expended on this area of difficulty. Here, we can imagine a role perhaps for something like crowdsourcing and majority opinion, to silence the minority who may be abusing the system.
4. CONCLUSION

We now reflect on continuing social challenges to be faced and provide some recommendations for the future. This paper has presented various technological solutions to the problem of cyberbullying. It is important to note that governments also have a role to play in the ongoing effort to combat this social problem and that legislation is still a part of the solution. But one message that resounds from our efforts on this topic is the following: computer scientists today have been trained to create effective technology for a host of applications. They tend to emerge with far less sensitivity to the social implications of the technology that they create. And yet, the skills that these people acquire can in fact be put to use in creating systems that will be able to make a difference with problems as endemic and as important as cyberbullying. The hope is that this paper has not only demonstrated some promise for making progress with the cyberbullying issue but has also helped to motivate computer scientists to become more engaged in efforts to address social implications of all kinds, through the use of technological solutions.

5. REFERENCES