Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele

Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability

Kathleen Van Royen [a,*], Karolien Poels [a], Walter Daelemans [b], Heidi Vandebosch [a]

[a] Media, ICT & Organisations and Society, Department of Communication Studies, University of Antwerp, Sint-Jacobstraat 2, 2000 Antwerp, Belgium
[b] CLiPS, Department of Linguistics, University of Antwerp, Lange Winkelstraat 40, 2000 Antwerp, Belgium

ARTICLE INFO

Article history:
Received 17 February 2014
Received in revised form 11 April 2014
Accepted 11 April 2014
Available online xxxx

Keywords:
Automatic monitoring
Cyberbullying
Adolescents
Social networking sites
Desirability

ABSTRACT

The automatic monitoring of cyberbullying on social networking sites has potential for signalling harmful messages, preventing these messages from remaining online and providing timely responses. Although technological advancements are made to optimise automatic cyberbullying detection systems, little is known about its desirability and requirements. Experts in the field of cyberbullying, as excellent sources of valuable insight into these issues, were solicited based on three open-ended questions relating to the desirability of automatic monitoring. Answers were examined through qualitative content analysis.

Of the 179 experts contacted, 50 (28%) responded. Most of these experts favoured automatic monitoring, but specified clear conditions under which such systems should be implemented, including effective follow-up strategies, protecting the adolescents' privacy and safeguarding their self-reliance.

Follow-up strategies should focus on preventing future cyberbullying and empowering the parties involved. The majority of respondents suggested priorities for detection, including threats and the misuse of pictures. Despite generally positive opinions, several experts harboured doubts regarding desirability and feasibility.

Appropriate follow-up strategies should be determined according to severity, and be tested for effectiveness. Future research should involve the views of adolescents and parents with regard to user desirability and prioritisation of cyberbullying detection, as well as views from social network providers.

1. Introduction

Cyberbullying is '*an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself*' (Smith et al., 2008) and can take multiple forms (e.g. threats, exclusion, name-calling) in different contexts (e.g. social networking sites, mobile phones) (Patchin and Hinduja, 2006; Willard, 2007a). Cyberbullying frequently occurs among adolescents on social networking sites (SNS) (Lenhart et al., 2011). It has been related to several emotional, psychological and physical problems (Hinduja and Patchin, 2007; Ybarra et al., 2006), as well as to poor academic performance (Tokunaga, 2010) and an increase in suicidal ideation (Hinduja and Patchin, 2010). Diverse impacts on victims have been observed, whether due to factors characterising

* Corresponding author. Tel.: +32 3 265 55 54; fax: +32 3 265 57 98.
E-mail address: kathleen.vanroyen@uantwerpen.be (K. Van Royen).

http://dx.doi.org/10.1016/j.tele.2014.04.002
0736-5853/© 2014 Elsevier Ltd. All rights reserved.

cyberbullying events or to differences in the resilience of the victims (Fenaughty and Harré, 2013; Ortega et al., 2012; Vandoninck et al., 2012; Ybarra et al., 2006).

Various strategies have been recommended for preventing and intervening in situations involving cyberbullying (Campbell, 2005; Cross et al., 2012; Perren et al., 2012). Examples of evidence-based, multi-component intervention programmes targeting adolescents, parents and schools include *Noncadiamointrappola!* (Palladino et al., 2012), *ConRed* (Ortega-Ruiz et al., 2012), and *Medienhelden* (Schultze-Krumbholz et al., 2012). These programmes, however, lack the technical resources required for a comprehensive approach to cyberbullying (Livingstone and Brake, 2010). SNS providers can play an important role in this regard by ensuring a safe environment, deleting harmful content and identifying perpetrators in severe cases (Vandebosch, 2014). In 2009, SNS providers active in Europe committed to ensuring the safety of young users by formulating the 'Safer Social Networking Principles', in consultation with the European Commission (EC Social Networking Task Force, 2009). Although they are not legally binding, these principles describe a number of safety strategies that can be employed on SNS, including the provision of educational messages and privacy protection, the empowerment of users and the installation of reporting mechanisms. One of these strategies involves having SNS providers monitor inappropriate content, thus allowing them to detect cyberbullying in an early stage, to take action and to reduce distress for victims (e.g. by preventing harmful content from remaining online). In current practice, SNS report to apply various mechanisms to a certain extent for reviewing their content in order to detect illegal or prohibited user-generated content, using human moderators or automated forms of monitoring (Staksrud and Lobe, 2010). Automatic monitoring seems particularly interesting, given the inherent impossibility of manually monitoring the millions of units of user-generated content every day on SNS in order to identify cyberbullying incidents.

To facilitate the process of screening large amounts of content, various initiatives are being taken to trace cyberbullying accurately and automatically (Dadvar et al., 2012, 2013; Dinakar et al., 2011). Automatic detection techniques use the same automatic text categorisation technology as proven applications such as spam filtering, topic detection, email routing etc. (Sebastiani, 2002). Although in principle, these detection models can be rule-based, and built by hand, machine learning approaches trained on sets of labelled examples dominate because of their ease of use, accuracy and efficiency. Obtaining this labelled data is expensive and time-consuming, but can be alleviated by using semi-supervised learning techniques which minimise the need for manual labelling (Delort et al., 2011). Given the difficulty of detecting cyberbullying compared to simpler types of unwanted content such as racist language or spam, more complex document representations are used and additional information about victims and bullies. For example, instead of only using words and emoticons expressing insults, profanity, and typical cyberbullying words, machine-learning models for cyberbullying can also take into account gender and personality of the participants in a potential cyberbullying event. This information can be automatically determined as well (Schwartz et al., 2013). Although development of automatic cyberbullying detection technology is in its early stages, and often with relatively low precision, it is nevertheless already useful by making the task of the human moderators easier. By focusing on the easier task of high recall (minimising the chance of false negatives), at the cost of high precision, the number of cases moderators have to check manually is significantly reduced.

Currently, studies on automatic cyberbullying detection are focusing mainly on its technological feasibility by optimising the accuracy of detection. In addition, insight into its desirability might be of equal importance in the decision to implement automatic monitoring. Concepts of feasibility and desirability are central to goal-setting in human decision making (Atkinson, 1957; Gollwitzer, 1990). Feasibility is being operationalized as the likelihood of attaining a goal, whereas desirability refers to the degree of the expected value, attractiveness or importance of the goal (Gollwitzer and Moskowitz, 1996; Gollwitzer, 1990). The attitude toward an action (its expected value) and the perceived controllability of this-action (its feasibility) conjointly determine whether an action is being executed (Ajzen, 1985). In a similar vein, both feasibility and desirability should be assessed for an optimal implementation of innovative technologies. In the current study, looking at automatic detection of cyberbullying as innovative technology, desirability will be operationalized as "the attitude on automatic monitoring of cyberbullying". To date, no research has been conducted on this issue, which calls for identifying the views of various stakeholders (e.g. adolescent SNS users, their parents, schools, cyberbullying experts). For this study we solicited views of experts in the field of cyberbullying. They can provide valuable insight in priorities and follow-up strategies, as they are familiar with the phenomenon of cyberbullying, as well as its context and impact. Moreover, their perception on the feasibility of recognising cyberbullying can be informative, as well as on requirements for the system in order to be desirable for its direct stakeholders.

In addition, it will be essential to know whether adolescents agree with the user conditions involved in such systems, as well as the forms of cyberbullying that they would like such systems to be able to detect. Understanding the attitudes and expectations of users with regard to safety measures on SNS is extremely important, as demonstrated by the reactions of users to changes in the features of Facebook, which reflected a widespread concern with privacy (Hoadley et al., 2010). The views of parents should also be considered, as automatic monitoring systems could be provided as features to be installed on home computers. It is therefore important to understand how such detection systems could affect the ways in which parents perceive safety in the context of SNS. Moreover, schools must be involved in assessing the desirability and informing the development of automatic monitoring systems as they are considered important actors in anti-cyberbullying initiatives (Vandebosch, 2014).

Finally, an automatic monitoring system should be developed in concert with SNS providers, who must ultimately adopt and implement monitoring systems, and consequently will be required to adjust and automate their current monitoring methods.

Please cite this article in press as: Van Royen, K., et al. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telemat. Informat. (2014), http://dx.doi.org/10.1016/j.tele.2014.04.002

The first phase of this study involved soliciting the views of experts regarding (1) the desirability of and requirements for automatic monitoring of cyberbullying on SNS, (2) their perceptions regarding the need for priorities in detection and (3) effective strategies for following up on cases after they have been detected.

2. Methods

Experts (*N* = 179) involved in research on cyberbullying and its prevention were invited to participate in this study. The experts included academic researchers (e.g. authors of recent publications) and members of international network initiatives targeted towards cyberbullying, including 'COST action IS0801 on cyberbullying' and the 'International Cyber Bullying Think Tank'. People involved in activities focused on prevention and awareness regarding cyberbullying or digital safety in Flanders and the Netherlands were contacted as well. A snowball method was used to identify additional experts through referral by other experts in the abovementioned fields. Experts were contacted by a personalised email with three open-ended questions concerning the desirability and prioritisation of automatic cyberbullying detection on SNS and what should happen after detection.

In order to avoid influencing their opinions, respondents were provided with neutrally formulated information on the purpose of the study, along with an objective description of automatic monitoring of cyberbullying on SNS. The first question assessed their opinions regarding the desirability of detection systems. All respondents completed this question. The next two questions addressed the need for priorities and various response strategies. They were to be answered only by those whose responses to the first question indicated a relative favourable attitude towards detection systems (39/50). Answers were examined through qualitative content analysis and thereby following a systematic way for describing the meaning of qualitative material (Schreier, 2012). '*Qualitative content analysis goes beyond merely counting words or extracting objective content from texts to examine meanings, themes and patterns that may be manifest or latent in a particular text*' (Zhang and Wildemuth, 2009). The meaning of the data was described systematically by classifying all data into the categories of a coding frame (Schreier, 2012). The coding frame consisted of dimensions derived from the research questions (e.g. 'desirability of automatic detection'; 'suggested priorities') generated in a data-driven way. Sub-categories were created within these dimensions, according to the responses (e.g. 'positive opinion on desirability'; 'requirements for automatic monitoring'), and the data were structured into these sub-categories.

In the following section, the results are presented according to the structure of dimensions and illustrated with quotations.

3. Results

Of the 179 experts contacted, 50 (28%) ultimately responded. The average age of the respondents was 45 years, with a gender distribution of 30% male and 70% female. Most of the respondents were involved in academic activities or research on cyberbullying. Others were involved with training and prevention initiatives on cyberbullying or digital safety in general, and some were educational or clinical psychologists.

3.1. Desirability of automatic monitoring of cyberbullying on social network sites

Most of the respondents expressed positive attitudes towards systems for automatically detecting cases of cyberbullying on SNS, although many specified conditions for the operationalisation of such systems, and several questioned their feasibility. Reasons motivating the positive attitudes expressed include the following: some adolescents lack the ability to judge or be aware that something is cyberbullying; many victims do not ask for help, and young and vulnerable users are in particular need of protection; parents and other educators lack the resources needed in order to react; and many SNS are not moderated. Other reasons included the possibility of deleting content in order to prevent messages from being spread. Moreover, some respondents noted that automatic detection could create a normative environment and increase awareness and education.

> *Especially because statistics show that young people who are facing this behaviour on the internet almost never ask for help, not from a parent, and not from a professional (like a teacher).*

> *Cyberbullying is very harmful, and the impact can be exacerbated by the ability to distribute posts and information widely, causing additional harm.*

Many experts indicated conditions for automatic detection. One condition advocated by many of the experts concerns effective follow-up strategies. Some respondents suggested a grading system, which would allow the classification of cyberbullying incidents according to their assessed severity and the needs of the parties involved in order to determine the appropriate response. A second condition proposed by the experts involves measures for protecting the privacy of adolescents by informing them about the monitoring system and its associated benefits. Other ideas included offering the system as optional for users (e.g. schools and other institutions).

> *Whether cyberbullying detection is desirable depends on what happens when cyberbullying is detected (and possibly verified by a human moderator).*

> *In my opinion, we should basically ensure that social network users are informed about what automatic detection is and how the device is operated.*

Some of the experts considered the automatic detection of cyberbullying undesirable. The main reasons articulated were ethical concerns regarding the violation of user privacy and serious doubts about the technological feasibility of developing a system that would be sensitive enough. They argued that the complex nature of certain cyberbullying cases and the subjective interpretation of cyberbullying would impair the accuracy of detection. They also noted that there is currently no consensus regarding the definition of cyberbullying.

> *No, I am not in favour of 'Big Brother' situations that involve spying on everything taking place online. Moreover, I believe that an automatic system would be fallible. To what extent would it filter out funny and teasing messages? To what extent can privacy be respected if social network sites start checking all of their content?*

Another concern involves the overprotection of adolescents, which they argued would decrease their ability to cope with distressing events.

> *Here lurks the risk of overprotection, excessive control, along with a decline in the resistance and resilience of adolescents.*

It was also argued that it would be better to invest in the education and empowerment of adolescents, as well as in the optimisation of reporting mechanisms, the awareness of report buttons, accompanied by the use of trained moderators to react more efficiently to reports of cyberbullying. Investment in reporting mechanisms was also suggested by those who were relatively in favour of automatic detection systems, as in situations that would be impossible to detect due to their complexity or invisibility.

Some experts also suggested the application of automatic warnings before uploading or writing inappropriate content. Other concerns included the possible unwillingness of social network providers to adopt and integrate such tool into their systems, the overload of detected content to be assessed and the possibility that adolescents might find other ways or platforms in which to bully each other.

Many experts, whether agreeing or disagreeing with automatic monitoring, in a way expressed the importance of respecting the autonomy of the adolescents, either to argument against automatic monitoring, either as a condition to be met in the monitoring system.

3.2. Prioritisation of types of cyberbullying for automatic monitoring

The majority of experts proposed priorities for detection. Most argued that detection should be focused on cases with the most dramatic consequences. The list of priorities included (1) threats involving physical assault or violence; (2) the misuse of pictures or videos of a pornographic, sexual or embarrassing nature; (3) cases demonstrating signs of suicidal ideation by victims; (4) hate speech (e.g. racism, homophobia); (5) commands to commit suicide; and (6) hate pages and fake profiles. Other criteria for prioritisation included frequently recurring cyberbullying, the extent of the event (e.g. number of bullies and bystanders involved; visibility to others; amount of offensive content), sexualised cyberbullying, defamation and personal denigration and whether the event is directed towards vulnerable people (e.g. young children).

The experts were also asked whether it might be less important to detect certain types of cyberbullying. In addition to indicating priorities for detection, many respondents stated that other cases should not be neglected and that all cases should eventually be considered and assessed. They argued that each case of cyberbullying has the potential to cause damage and that their severity and impact are highly dependent upon the vulnerability of individual victims.

> *All cyberbullying has the potential to cause damage to young people's wellbeing, and it should be prevented/addressed effectively*.

> *It is obvious that there are gradations in bullying. However before this can be evaluated, the situation should be thoroughly assessed. This is only possible if all forms of cyberbullying are taken into account. Situations that appear ostensibly innocent, could be alarming upon closer inspection.*

It was noted, however, that some events simply do not lend themselves to detection (e.g. name-calling, due to differences in culture and interpretation).

In contrast, some respondents stated that it would be impossible to assess all content and that it is important to respect the experimental nature of adolescents' behaviour. They mentioned cases that would be less important to detect, including incidents of teasing, gossiping, name-calling, single events and cases in which the victims are able to react on their own (e.g. by reporting incidents of direct harassing messages).

> *All cases in which people can react on their own or in which the network of friends reacts in defence. It would be wrong to replace this positive self-reliance with an acquired dependent attitude that involves expecting a reaction from the SNS.*

> *Prioritisation is important [...] Focusing on too many themes is unwise. Users will report it themselves, if they are being hacked or being threatened for instance. If you identify too much, it will yield more data that should be followed up. Providers won't savour this.*

To further determine priorities, it will be important to include adolescents' views, as indicated by this expert.

> *In my opinion there are no forms which are really less important. But in the opinion of youth, we see that disturbing talk in chat rooms are not so harmful.*

3.3. Follow-up on monitoring

Experts were asked to indicate appropriate responses to cases in which cyberbullying content has been detected. Follow-up was considered the most important phase of the monitoring process.

Several respondents suggested that it is important to employ trained people who would carefully appraise the detected case according to severity in a neutral way and who would have access to the appropriate resources with which to react accordingly.

> *Detecting cyberbullying is a first important step, however the next step will be even more important. The situation must be evaluated by people with expertise who at the same time are capable to act if necessary*.

According to the experts, removing or blocking the content in question should be accompanied by actions directed towards both the aggressor and the victim. The most frequently mentioned response to aggressors involved the ban of their profiles (mostly temporary). Many of the experts specified that such bans should be imposed only after repeated incidents, with single events followed by a warning. Several of the experts, however, felt that it would be impossible to enforce a sanction like 'banning' an adolescent from SNS. Further suggestions included efforts to achieve behavioural change on the part of the aggressors (e.g. educational games, awareness videos or temporary warnings generated whenever an aggressor is about to post a new message).

Responses towards victims included the provision of advice on coping and support. Educational efforts for future situations were also suggested, including empowering both victims and bystanders to seek help or report offensive content.

> *The primary responsibility of providers is to ban any harmful material from their sites. They should therefore remove any such material as soon as possible (photos, videos, comments, etc.). Furthermore, it would be desirable to refer victims to appropriate institutions for further help or counselling, if necessary. And it might be helpful for perpetrators to experience some negative consequences – at least if they are repeated bullies – such as the deletion of their profiles.*

Several of the respondents proposed that the first step after detection should involve contacting victims to ask whether they considered the incidents as harassing.

Some experts expressed doubts regarding the role of SNS providers in follow-up efforts, as they would not be qualified to do so. For this reason, several respondents proposed collaboration with and referral to actors in 'real life'. If illegal acts were involved, the appropriate institutions or authorities should be contacted. Victims should be offered referral to professional help, and other actors (e.g. parents and teachers) could be contacted for support.

Several respondents recommended identifying responses that do and do not work, in addition to examining the situations in which who should be contacted and how.

4. Discussion

This study identifies the views of cyberbullying experts with regard to the desirability of and requirements for automatic cyberbullying detection systems for SNS. To our knowledge, it is the first study to address these issues. The results can be used to inform the technical development process, to establish priorities for detection and to develop an appropriate follow-up system. It is important to note, however, that this study is based solely on the views of a limited sample of cyberbullying experts. More stakeholders should be involved before any definitive conclusions are drawn. It is also important to interpret these results with caution, as experts who were not in favour of automatic monitoring might have been more inclined to respond to this study or, conversely, those with a supportive attitude might have been more likely to express their opinions.

4.1. Risk protection or self-reliance?

Most of the cyberbullying experts were in favour of the automatic monitoring of cyberbullying on SNS, although they specified conditions that should be met, including the installation of measures to protect the privacy of adolescents. In a similar vein, the concerns expressed by the experts who were not in favour of automatic monitoring involved ethical issues relating to the potential invasion of adolescents' privacy. According to the United Nations Convention on the Rights of the Child (UNCRC), children have the right to be protected from harm. However, the UNCRC also states that children have a right for provision of internet resources and freedom of expression (participation rights) (Livingstone and O'Neill, 2014). Therefore such concerns must be considered throughout the development of any automatic detection system. The extent to which these concerns are grounded could be addressed further by soliciting the views of adolescents regarding automatic monitoring. Young users appear to be more concerned about social privacy than they are about institutional privacy, indicating that they are more likely to focus on controlling their information with respect to other people than with respect to Facebook or other corporations (Raynes-Goldie, 2010). If this is the case, they might be more likely to favour monitoring by SNS than by their parents. In particular, studies have shown that adolescents are more concerned about maintaining their online privacy from their parents (Livingstone and Bober, 2006) and that they tend to perceive online monitoring by parents as reflecting a lack of respect for their ability to make responsible decisions and choices (Media Awareness Network, 2004). In this study, the need to safeguard the independence of adolescents throughout the detection process emerged as an important issue. Several of the experts expressed the concern that automatic monitoring might cause adolescents to become overprotected and lose their ability to cope with cyberbullying. This concern relates to the trend towards a culture of fear dominated by surveillance (Marx and Steeves, 2010), and it raises questions regarding how to achieve a proper balance between self-reliance and risk protection and between free expression and surveillance. Moreover, some experts suggested that situations to which adolescents are able to respond on their own should not be detected. These results illustrate the importance of ensuring the self-reliance of adolescents (e.g. by considering their opinions on what should be detected and by identifying whether they perceive automatic monitoring as a threat to their autonomy). Another plausible solution for preserving the autonomy of adolescents might involve the use of ‘reflective user interfaces’, such as notifications that urge users to reflect in anticipation on posting potentially harmful content on SNS (Dinakar et al., 2012).

In addition to loss of autonomy, another potential drawback of automatic monitoring is the creation of a false sense of security for parents and adolescents. Therefore it should be emphasised that automatic monitoring of cyberbullying on SNS would serve as one protective mechanism in addition to prevention and awareness raising initiatives.

4.2. Balancing interests

These results provide no insight into the views of social network providers. To date, surveillance in the context of Web 2.0 remains a relatively unstudied area (Fuchs, 2010). The studies that have been conducted tend to focus on horizontal peer surveillance (Albrechtslund, 2008; Marwick, 2012) or vertical surveillance with regard to economic interests (Andrejevic, 2011; Cohen, 2008; Fuchs, 2011b, 2012). The processing of user data for positive ends, referred to as ‘monitoring’ by Fuchs (2011a), remains unexplored. Web 2.0 surveillance studies exploring perceptions focus on users (Farinosi, 2011; Fuchs, 2010; Jansson, 2012; Taddicken, 2011). However, particularly, the field of monitoring would benefit from views of social network providers into their desirability, as they are the ones who must adopt such systems.

Despite the efforts made by SNS to guarantee the online safety of their users (EC Social Networking Task Force, 2009), such providers are primarily oriented towards making a profit in order to serve their commercial interests (Langlois et al., 2009), using data for advertising purposes. The extent to which social network providers will be inclined to implement automatic monitoring tools for cyberbullying into their platforms thus remains to be seen, particularly given the crucial requirement of organising appropriate follow-up strategies, as emphasised by the experts in this study. Commitment to protecting children from harm by SNS could be increased by referring to their corporate social responsibility. European policy makers can contribute in this regard by aligning strategies to promote corporate social responsibility of SNS providers. Thus far, it is a positive sign that the majority of SNS have committed to the Safer Networking Principles and have generally demonstrated good to fair compliance with these principles (Donoso, 2011; Staksrud and Lobe, 2010). According to the latest self-declaration reports of SNS, their compliance is rather satisfactory to very satisfactory with regard to their response to reports from users and reviewing illegal or prohibited content and conduct (Donoso, 2011).

4.3. Detection focusing on threats and the misuse of pictures

Most of the respondents suggested priorities for the focus of detection tools, mainly based on objective criteria (e.g. threats) but also on individual outcomes (e.g. signs of suicidal ideation). Given that studies have demonstrated that some cases of cyberbullying are likely to cause more distress to victimised adolescents than others are (Fenaughty and Harré, 2013; Staude-Müller et al., 2012; Ybarra et al., 2006), it might be a good strategy to align certain priorities. The experts mentioned forms of cyberbullying involving threats and the misuse of pictures as priorities far more frequently than they did other forms, and such incidents are consistently bounded by legal provisions and the Terms of Service of SNS (Lievens, 2012). In addition, empirical evidence has demonstrated that the misuse of sexually explicit or embarrassing pictures causes more distress for victims than do other forms of cyberbullying (Slonje and Smith, 2008). Adopting a cross-media detection approach therefore is essential, by reviewing both textual and visual posts on SNS to detect cyberbullying.

Despite the relevance of these priorities, it is vital to learn more about the perceptions of adolescents with regard to the cases to be detected. Additional research must be conducted in order to align the priorities of the various parties involved.

Even though most experts stated priorities for detection, they disagreed on whether some cases deserve less attention. Many experts mentioned that some of the less important cases should not be ignored, whereas others argued that it is less important to focus on some forms of cyberbullying. These results suggest the challenge of accurately defining exactly what constitutes ‘real’ bullying and what should be considered part of adolescents’ experimental behaviour. Apart from the accurate alignment of which cyberbullying cases should be detected, another challenge involves the conceptual understanding of cyberbullying, as addressed in a considerable number of theoretical studies (Gradinger et al., 2010; Langos, 2012; Menesini and Nocentini, 2009). Considerable caution should be exercised when labelling an interaction as cyberbullying. In a similar vein, Guldberg (2009) notes an increase in the tendency to label forms of interaction as 'bullying', with the result that increasing numbers of children are being labelled as either 'bullies' or 'victims'. In the process of developing and using any automatic monitoring system, it is advisable to take care to avoid succumbing to the contemporary obsession with bullying, in which the behaviour of adolescents is viewed with the same seriousness as that of adults (Guldberg, 2009). Marwick and boyd describe (2011) that teens, especially girls, are engaging often in 'drama' on SNS, including forms of gossip, arguing, ostracization and name-calling, which closely resembles bullying. Teens report this 'drama' as normal part of their lives. This poses the challenge to distinct forms of 'drama' from bullying by the monitoring system.

Similarly, several experts harboured doubts concerning the feasibility of an automatic detection system for cyberbullying, due to the complexity and subjectivity of the phenomenon. The detection of cyberbullying is far more complex than is the detection of spam or other internet abuses, given its highly personalised and contextual nature (Lieberman et al., 2011; Yin et al., 2009). Despite this complexity, however, researchers studying detection are convinced of its technological feasibility. They argue that cyberbullying occurs around a very limited number of topics, such as race, physical appearance and sexuality (Dinakar et al., 2012; Mishna et al., 2010). In combination with information on the tone of the message, cues relating to these topics could enhance the identification of many messages that could potentially constitute cyberbullying (Lieberman et al., 2011). Additional studies examining whether the incorporation of user context and other variables allow for more accurate detection would be particularly useful in this regard (Dadvar et al., 2013), as would commonsense reasoning (Dinakar et al., 2012) and sentiment analysis (Nahar et al., 2012).

4.4. Follow-up strategies

The experts participating in this study identified effective follow-up as the most important condition for the automatic monitoring process. Suggestions for effective responses ranged from technical solutions by SNS providers (e.g. removing offensive content or temporarily banning aggressors) to intervention strategies aimed at both aggressors and victims. The latter solutions included the provision of educational messages and advice, as well as referral to professionals, parents, schools or other authorities for further follow-up. A cyberbullying detection system that implements appropriate follow-up could enhance both awareness and prevention, through possibly empowering victims to cope with problems in the future and making aggressors aware of the inappropriate character of their behaviour. The autonomy of adolescents is emphasised by experts in this regard as well, with arguing for the empowerment of victims and bystanders in order to cope with potential future incidents. Suggested response and prevention actions stress the need to empower victims of cyberbullying (Perren et al., 2012; Willard, 2007b).

Measures should be included to respect the child's right for freedom of expression (Livingstone and O'Neill, 2014), for instance by contacting the victim before deleting content or any further steps, as was suggested by several experts.

Moreover, several of the respondents preferred to start by assessing the detected case and responding accordingly or to offer instrumental support through referral. To date little empirical evidence is available on the effectiveness of various responses towards cyberbullying (Perren et al., 2012). Further research is needed to examine ways of administering teams of trained experts for cases reported by users, as well as for cases identified through automatic detection. Health professionals and other experts would be well equipped to provide further support in establishing appropriate follow-up strategies.

4.5. Future steps

In conclusion, more reflection is essential on the balance between rights and interests of different actors. Therefore additional research is needed on the views of stakeholders such as adolescents, parents and SNS providers with regard to the automatic monitoring of cyberbullying on SNS.

The misuse of pictures being considered a high priority suggests strategies for automatic detection such as a cross-media detection approach focusing both on visual and textual cyberbullying. Furthermore, automatic warnings before uploading and other alternatives for automatic detection to allow for prevention of harm should be examined as well. Finally, a response grading system could be developed, through which cases could be classified according to assessments of severity and subsequently linked to appropriate follow-up measures. Each possible response should be tested for effectiveness, as well as for the feasibility of having it implemented by SNS providers. In addition, once automatic monitoring would be implemented, evaluation of the follow-up strategies being used will be very important. Even if automatic monitoring of cyberbullying were to prove undesirable for adolescents or other stakeholders, also for reported content by users it will be crucial to enable improved and appropriate responses to offensive content.

Acknowledgments

The authors thank all experts who participated in this study.

References

Ajzen, I., 1985. From intentions to actions: a theory of planned behavior. In: Kuhl, P.D.J., Beckmann, D.J. (Eds.), Action Control. Springer, Berlin Heidelberg, pp. 11–39.

Albrechtslund, A., 2008. Online social networking as participatory surveillance. First Monday 13 (3). Available from: <http://firstmonday.org/ojs/index.php/fm/article/view/2142/1949/>.

Andrejevic, M., 2011. Surveillance and alienation in the online economy. Surveill. Soc. 8 (3), 278–287.

Atkinson, J., 1957. Motivational determinants of risk-taking behavior. Psychol. Rev. 64, 359–372.

Campbell, M.A., 2005. Cyber bullying: an old problem in a new guise? Aust. J. Guid. Couns. 15 (1), 68–76.

Cohen, N., 2008. The valorization of surveillance: towards a political economy of Facebook. Democr. Commun. 22 (1), 5–22.

Cross, D., Li, Q., Smith, P.K., Monks, H., 2012. Understanding and preventing cyberbullying. Where have we been and where should we be going? In: Li, Q., Cross, D., Smith, P.K. (Eds.), Cyberbullying in the Global Playground. Research from International Perspectives. Wiley-Blackwell, West Sussex, pp. 287–305.

Dadvar, M., Ordelman, R., de Jong, F., Trieschnigg, D., 2012. Towards user modelling in the combat against cyberbullying. In: Proceedings of the17th International Conference on Applications of Natural Language to Information System, Groningen, The Netherlands, pp. 277–283.

Dadvar, M., Trieschnigg, D., Ordelman, R., de Jong, F., 2013. Improving cyberbullying detection with user context. In: Proceedings of the 35th European Conference on IR Research, Berlin, vol. 7814, pp. 693–696.

Delort, J.-Y., Arunasalam, B., Paris, C., 2011. Automatic moderation of online discussion sites. Int. J. Electron. Comm. 15 (3), 9–30.

Dinakar, K., Reichart, R., Lieberman, H., 2011. Modeling the detection of textual cyberbullying. In: Presented at the International AAAI Conference on Weblogs and Social Media, Barcelona, Spain. Available from: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3841/4384>.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., Picard, R., 2012. Commonsense reasoning for detection, prevention and mitigation of cyberbullying. ACM Trans. Interact. Intell. Syst. 2 (3).

Donoso, V. 2011. Assessment of the Implementation of the Safer Social Networking Principles for the EU on 14 Websites: Summary Report (p. 79). European Commission, Luxembourg. Available from: <http://ec.europa.eu/information_society/activities/social_networking/docs/final_report_11/part_one.pdf>.

EC Social Networking Task Force, 2009. Safer Social Networking Principles for the EU. European Commission, Luxembourg. Available from: https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/sn_principles.pdf.

Farinosi, M., 2011. Deconstructing Bentham's panopticon: the new metaphors of surveillance in the Web 2.0 environment. tripleC 9 (1), 62–76.

Fenaughty, J., Harré, N., 2013. Factors associated with distressing electronic harassment and cyberbullying. Comput. Hum. Behav. 29, 803–811.

Fuchs, C., 2010. StudiVZ. Social networking in the surveillance society. Ethics Inf. Technol. 12 (2), 171–185.

Fuchs, C., 2011a. New media, Web 2.0 and surveillance. Sociol. Compass 5 (2), 134–147.

Fuchs, C., 2011b. Web 2.0, prosumption, and surveillance. Surveill. Soc. 8 (3), 288–309.

Fuchs, C., 2012. The political economy of privacy on Facebook. Telev. New Media 13, 139–159.

Gollwitzer, P., 1990. Action phases and mind-sets. Handbook of Motivation and Cognition. Guilford Press, New York, pp. 53–92.

Gollwitzer, P., Moskowitz, G., 1996. Goal effects on action and cognition. In: Social Psychology: Handbook of Basic Principles. Guilford Press, New York, pp. 361–399.

Gradinger, P., Strohmeier, D., Spiel, C., 2010. Definition and measurement of cyberbullying. Cyberpsychol. J. Psychosoc. Res. Cyberspace 4 (2). Available from: <http://cyberpsychology.eu/view.php?cisloclanku=2010112301>.

Guldberg, H., 2009. Reclaiming Childhood. Freedom and Play in an Age of Fear. Routledge, Abingdon, Oxon, pp. 92–110 (Vol. The Bullying Bandwagon).

Hinduja, S., Patchin, J., 2007. Offline consequences of online victimization. J. Sch. Violence 6 (3), 89–112.

Hinduja, S., Patchin, J.W., 2010. Bullying, cyberbullying, and suicide. Arch. Suicide Res. 14 (3), 206–221. http://dx.doi.org/10.1080/13811118.2010.494133.

Hoadley, C., Xu, H., Lee, J., Rosson, M., 2010. Privacy as information access and illusory control: the case of the Facebook News Feed privacy outcry. Electron. Commer. Res. Appl. 9, 50–60.

Jansson, A., 2012. Perceptions of surveillance: reflexivity and trust in a mediatized world (the case of Sweden). Eur. J. Commun. 27 (4), 410–427.

Langlois, G., McKelvey, F., Elmer, G., Werbin, K., 2009. Mapping commercial Web 2.0 worlds: towards a new critical ontogenesis. Fibreculture J. (14). Available from: <http://fourteen.fibreculturejournal.org/fcj-095-mapping-commercial-web-2-0-worlds-towards-a-new-critical-ontogenesis/>.

Langos, C., 2012. Cyberbullying: the challenge to define. Cyberpsychol. Behav. Soc. Netw. 15 (6), 285–289. http://dx.doi.org/10.1089/cyber.2011.0588.

Lenhart, A., Madden, M., Smith, A., Purcell, K., Zickuhr, K., Rainie, L., 2011. Teens, Kindness and Cruelty on Social Network Sites. How American Teens Navigate the New World of "Digital Citizenship". Pew Research Center's Internet & American Life Project, Washington, D.C.

Lieberman, H., Dinakar, K., Jones, B., 2011. Let's gang up on cyberbullying. Computer 44 (9), 93–96.

Lievens, E., 2012. Bullying and Sexting in Social Networks from a Legal Perspective: Between Enforcement and Empowerment (ICRI Working Paper 7/2012). Interdisciplinary Centre for Law and ICT, K.U. Leuven. Available from: <https://www.law.kuleuven.be/icri/ssrnpapers/35ICRI_Working_Paper_7_2012.pdf>.

Livingstone, S., Bober, M., 2006. Regulating the internet at home: contrasting the perspectives of children and parents. In: Digital Generations: Children, Young People and New Media. Lawrence Erlbaum Associates, Mahwah, New Jersey, pp. 93–113.

Livingstone, S., Brake, D., 2010. On the rapid rise of social networking sites: new findings and policy implications. Child. Soc. 24 (1), 75–83.

Livingstone, S., O'Neill, B., 2014. Chapter 2 children's rights online: challenges, dilemmas and emerging directions. In: Minding Minors Wandering the Web: Regulating Online Child Safety. Asser Press, The Hague, pp. 19–38.

Marwick, A., 2012. The public domain: social surveillance in everyday life. Surveill. Soc. 9 (4), 378–393.

Marwick, A.E., boyd, Danah, 2011. The Drama! Teen Conflict, Gossip, and Bullying in Networked Publics (SSRN Scholarly Paper No. ID 1926349). Social Science Research Network, Rochester, NY. Available from: <http://papers.ssrn.com/abstract=1926349>.

Marx, G., Steeves, V., 2010. From the beginning: children as subjects and agents of surveillance. Surveillance & Society 7 (3/4), 192–230.

Media Awareness Network, 2004. Young Canadians in a Wired World: Phase II. Focus Groups. Media Awareness Network, Ota.

Menesini, E., Nocentini, A., 2009. Cyberbullying definition and measurement: Some critical considerations. Z. Psychol./J. Psychol. 217 (4), 230–232.

Mishna, F., Cook, C., Gadalla, T., Daciuk, J., Solomon, S., 2010. Cyber bullying behaviors among middle and high school students. Am. J. Orthopsychiatry 80 (3), 362–374.

Nahar, V., Unankard, S., Li, X., Pang, C., 2012. Sentiment analysis for effective detection of cyber bullying. Lect. Notes Comput. Sci. 7235, 767–774.

Ortega, R., Elipe, P., Mora-Merchán, J.A., Genta, M.L., Brighi, A., Guarini, A., Smith, P.K., Thompson, F., Tippett, N., 2012. The emotional impact of bullying and cyberbullying on victims: a European cross-national study. Aggress. Behav. 38 (5), 342–356. http://dx.doi.org/10.1002/ab.21440.

Ortega-Ruiz, R., Del Rey, R., Casas, J.A., 2012. Knowing, building and living together on internet and social networks: the Conred cyberbullying prevention program. Int. J. Confl. Violence. 6 (2), 302–312.

Palladino, B.E., Nocentini, A., Menesini, E., 2012. Online and offline peer led models against bullying and cyberbullying. Psicothema 24 (4), 634–639.

Patchin, J.W., Hinduja, S., 2006. Bullies move beyond the schoolyard. A preliminary LOOK at cyberbullying. Youth Violence Juv. Justice 4 (2), 148–169.

Perren, S., Corcoran, L., Cowie, H., Dehue, F., Garcia, D., Mc Guckin, C., Sevcikova, A., Tsatsou, P., Völlink, T., 2012. Tackling cyberbullying: review of empirical evidence regarding successful responses by students, parents, and schools. Int. J. Confl. Violence 6 (2), 283–293.

Raynes-Goldie, K., 2010. Aliases, creeping, and wall cleaning: understanding privacy in the age of Facebook. First Monday 15 (1). Available from: <http://firstmonday.org/ojs/index.php/fm/article/view/2775/2432>.

Schreier, M., 2012. Qualitative Content Analysis in Practice. Sage Publications, Chennai, India.

Schultze-Krumbholz, A., Wölfer, R., Jäkel, A., Zagorscak, P., Scheithauer, H., 2012. Effective Prevention of Cyberbullying in Germany – The Medienhelden Program. Oral Presentation presented at the XXth ISRA World Meeting, Luxembourg.

Please cite this article in press as: Van Royen, K., et al. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telemat. Informat. (2014), http://dx.doi.org/10.1016/j.tele.2014.04.002

Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., Ungar, L.H., 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS One 8 (9), e73791. http://dx.doi.org/10.1371/journal.pone.0073791.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34 (1), 1–47.

Slonje, R., Smith, P.K., 2008. Cyberbullying: another main type of bullying? Scand. J. Psychol. 49 (2), 147–154.

Smith, P.K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., Tippet, N., 2008. Cyberbullying: its nature and impact in secondary school pupils. J. Child Psychol. Psychiatry 49 (4), 376–385.

Staksrud, E., Lobe, B., 2010. Evaluation of the Implementation of the Safer Social Networking Principles for the EU Part I: General Report. European Commission Safer Internet Programme, Luxembourg.

Staude-Müller, F., Hansen, B., Voss, M., 2012. How stressful is online victimization? Effects of victim's personality and properties of the incident. Eur. J. Develop. Psychol. 9 (2), 260–274.

Taddicken, M., 2011. Self-disclosure in the social web: exploring users' privacy and surveillance concerns via focus groups. In: Presented at the annual meeting of the International Communication Association, Boston.

Tokunaga, R.S., 2010. Following you home from school: a critical review and synthesis of research on cyberbullying victimization. Comput. Hum. Behav. 26 (3), 277–287.

Vandebosch, S., 2014. Chapter 14 addressing cyberbullying using a multi-stakeholder approach: the Flemish Case. In: Minding Minors Wandering the Web: Regulating Online Child Safety. Asser Press, The Hague, pp. 243–260.

Vandoninck, S., D' Haenens, L., Segers, K., 2012. Coping and resilience: children's responses to online risks. In: Children, Risk and Safety on the Internet. Research and Policy Challenges in Comparative Perspective. The Policy Press, Bristol, pp. 205–218.

Willard, N., 2007a. Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Research Press, Illinois.

Willard, N., 2007b. Response actions and options. In: Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Research Press, Illinois, pp. 141–153.

Ybarra, M.L., Mitchell, K.J., Wolak, J., Finkelhor, D., 2006. Examining characteristics and associated distress related to internet harassment: findings from the second youth internet safety survey. Pediatrics 118 (4), e1169–e1177.

Yin, D., Xue, Z., Hong, L., Davison, B., Kontostathis, A., Edwards, L., 2009. Detection of Harassment on Web 2.0. In: Presented at the Proceedings of the 1st Content Analysis in Web 2.0 Workshop, Madrid, Spain.

Zhang, Y., Wildemuth, B., 2009. Qualitative Analysis of Content. Applications of Social Research Methods to Questions in Information and Library Science. Libraries Unlimited, Westport, CT, pp. 308–319.

Please cite this article in press as: Van Royen, K., et al. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telemat. Informat. (2014), http://dx.doi.org/10.1016/j.tele.2014.04.002