Telematics and Informatics 32 (2015) 89–97

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 23 April 2014

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.