






Detection and Prevention of Bullying on Online Social Networks: The Combination of Textual, Visual and Cognitive

Carlos Silva , Ricardo Santos , and Ricardo Barbosa 

CIICESI - Center for Research and Innovation in Business Sciences
and Information Systems, School of Management and Technology,
Polytechnic Institute of Porto, Felgueiras, Portugal
{8120333, rjs, rmb}@estg.ipp.pt

Abstract. The adoption of online social platforms as a common space for the virtualisation of identities is also correlated with the replication of real-world social hazards in the virtual world. Bullying, or cyberbullying, is a very common practice among people nowadays, becoming much more present due to the increase of online time, especially in online social networks, and having more serious consequences among younger audiences. Related work includes the analysis and classification of textual characteristics that can be indicative of a bullying situation and even a visual analysis approach through the adoption of image recognition techniques. While agreeing that the combination of textual and visual analysis can help the identification of bullying practice, or the identification of bullies, we also believe that a part is missing. In this work, we propose a combination of textual and visual classification techniques, associated with a cognitive aspect that can help to identify possible bullies. Based on a previous model definition for a virtual social sensor, we propose the analysis of textual content present on online social networks, check the presence of people in multimedia content, and identification of the stakeholders on a possible bullying situation by identifying cognitive characteristics and similarities on the behaviours of possible bullies and/or victims. This identification of possible bullying scenario can help to address them before they occur or reach unmanageable proportions.

Keywords: Online social networks · Bullying · Virtual social sensor · Cognitive

1 Introduction

The high dependence on the computer and mobile devices, whether to professional or academic work, either for leisure or other activities such as reading news or consulting social networks, is strongly present on the Internet. The Internet is a great asset because it allows us to be connected to any part of the world in a fraction of a second, enabling us to access any information in real time, and it can override other methods that would take us much longer to achieve certain goals. However, it is not full of wonders, and

each user is constantly susceptible to the range of some attacks such as virus or phishing and is subject to believe in misleading information.

Children spend some time with the new technologies and accessing the Internet in an uncontrolled way by their parents or tutors may lead them to use it improperly. The visualization of content that they should not have access to, could make them a target for social hazards. Focusing on the usage of online social networks (Facebook, Twitter, Youtube, Instagram, just to name a few) we face a set of problems like lack of privacy, or even the possibility of being victims of insult or other harming situations, in order words, bullying.

Bullying is a complex social dynamic, motivated essentially by differences of the domain, social capital or culture [1]. The desire for dominance, acquisition and maintenance of social capital are the main factors of motivation for the initiation and prolongation of the practice of bullying. For example, the lack of social capital by the victims may impede them from getting a better social position or the capacity to acquiring a specific thing, which may lead to contempt from the others. In addition, the denomination used by aggressors, also known as bullies, to subjugate the victims, results in an intense humiliation that has negative effects on such people, like anger and depression.

Cyberbullying, like traditional bullying, has a profound negative impact on the victim, especially when dealing with children and young people, suffering significantly in one emotional and psychological way, even with some cases ending in tragic suicides. So, cyberbullying can be described as: when the Internet, mobile phones or other devices are used to send text or images that can hurt, humiliate or embarrass other people, and it is a more constant version of the traditional bullying [2].

Unlike spam, this kind of attack is more personal, varied and contextual [3]. The images published by an individual in a social network, the type of content shared, the links comments and the possibility of easy exchange of messages with any other user, allow the practice of bullying to be more frequent and constant than ever, and it is a danger that must be considered in our society.

In this work, we aim to address this problem by combining three main types of classification techniques: textual; visual; and cognitive. Taking as base a previously defined virtual social sensor model, we enhance its capabilities by providing new modules that can help the identification of bullying situations, correctly identification of bullies, and the possibility to identify potential bullies, or bullying situations, through the analysis of personality traits (that are directly responsible for our behaviour).

This work is divided as follows: Sect. 2 describes the main problems that arise on the Internet, and introduces the topic of online bullying, or cyberbullying; Sect. 3 is dedicated to an overview look of the work already developed in this subject, describing the approaches and technologies used for the development and implementation. In Sect. 4 we describe our solution proposal, which includes the combination of the three main characteristics (textual, visual, and cognitive) as an approach to correctly identify bullies, bullying, and other social hazards. This work ends with a conclusion and definition of the future work.

2 Internet as a Threat

Our constant need to be connected with the world, either for short periodic interactions, or for prolonged activities like viewing the news, email, interact in social platforms, or talk with a customer or partner in order to conduct deals, makes unbelievable to live without the Internet. The Internet allows us to perform an infinite number of tasks and is intended to simplify the life of those who use it, becoming an added value for society. However, behind all those benefits, its usage hides a negative side.

A study [4] analysed children behaviour on the Internet across several countries, and from all risks present across the web for children, (between 11 and 16 years old) the authors identified that 5% of the respondents (about 3500 in total) already suffered bullying, and another 5% received messages of a sexual nature, however, only 3% felt bothered by that situation. Meeting new people was reported by 11% of the children, and none of them felt uncomfortable with that. The contact with images that contained nudity or pornographic content was experienced by 27%, and other types of content related to hate, self-harming, anorexia and drugs were reached by 10%. The contact with the technologies and the Internet begins by playing games, which in many cases require connection to the Internet, and may allow connection to other unknown players in any part of the planet. Then, they need to use the web for research to the accomplishment of school works, followed by joining social networks, mainly Facebook, Twitter and Instagram. This data shows that children begin to have access to technology in a prematurely way, coinciding mostly with the period in which they are attending the first cycle of basic education.

Recently, big organisations like Google or Facebook, are taking more attention to these situations to implement web-based security programs, against several types of hazards. This kind of actions would be necessary to help avoid facing these situations, despite the existence of advisor programs that alert people for some of these cases that are broadcasted mainly in the schools and in the traditional media (like television or newspaper). The introduction of automatic mechanisms may prevent the occurrence of some of these problems or control them to avoid scaling up without control.

2.1 Bullying Is also Cyberbullying

On the Internet, it is common to have a reflection of some social problems that exist in the real world, like bullying. Bullying is the process of threatening or assaulting an individual or a group of individuals towards others, usually related to some characteristic of their lives, as their culture, and is more common among young people.

On the virtual world, the practice of bullying, or cyberbullying, is performed by repeated psychological violence acts, practised by one youth or groups of young people over another, using technologies, either using Internet applications or directly through text messages and telephone calls. Unlike traditional bullying, bullying through technologies do not lead to physical contact, unless those involved met on daily bases, such as at school, where the situation can get worse. However, given that young people spend a lot of time with their technological devices, especially for checking and updating of their social networks, this practice of violence becomes more constant, harder to identify, and more conducive to humiliation with a higher number of people reached.

McClowry et al. [5] divide bullying into two types: direct, entails blatant attacks on a targeted young person; indirect, involves communication with others about the targeted individual (spreading harmful rumours). The authors also refer out that bullying can be physical, verbal or relational (excluding someone, denying friendship) and may involve property damage. Boys tend to do more direct bullying behaviours, while girls are more involved in acts related to indirect bullying.

Hee et al. [6] describe three main roles in a bullying scenario, the bully, the victim and bystanders. The bullies are those ones who intend to attack or threat someone, the victims correspond to the people who suffer bullying from others and bystanders are those who also view the post and sometimes interact in it. They can show up to support the victim and mitigate the negative effects caused by the bullying or can be an active part in the bullying, joining the attacker and making the situation go worse or can simply ignore the evidence and keep scrolling on the news feed.

But why would anyone take the initiative to attack another? In many cases, the bully has been a victim before, making him a more furious and aggressive person causing him to use that anger on someone, feels solitude and needs attention, has problems at home, or has low self-esteem and to feel better tries to degrade others. The bully may also want to have more popularity and attacks people he feels jealous of, can have a big ego, may feel superior to others, and has a security group in most of the cases to feel safe if someone riposte to his attacks [7]. Often, these attacks are linked to sensitive topics such as race and culture, sexuality, intelligence, physical appearance, and aspects that people cannot change about themselves [2].

Sometimes the target of offensive messages on the Internet is also the one who writes them, sending it to themselves, under a pseudonym. The motives for that vary, from young people who do it as a form of fun, to people who want to test the reaction of some friends, or cases of depressed individuals who want to make themselves feel even worse. This behaviour is more prevalent in adolescents who do not identify themselves as heterosexuals and in people who had been victims of bullying in the past. Boys are also more likely to send offensive messages to themselves, usually as a joke to get the attention of friends, or love interests [8].

Cyberbullying is also much more likely to be done by someone the victim knows well. Children are seven times more likely to be attacked by current or former friends or romantic interests than by any stranger. More than a third of adults harassed on the web do not know the person who is harassing them, and just less than a third are harassed by people who hide their identities. Homosexual students are more likely to be victims of these acts, as are non-white students. Girls are 2.6 times more likely to be victims than boys, and woman are 2 times more likely to be harassed online. In some cases, through fake accounts, the attacker tries to impersonate the victim, publishing content to humiliate that person [9]. Bullies normally post less, participate in fewer online communities, and are less popular than normal users [10].

This kind of attack may look easy to dismiss, however, the perpetrator blackmails the victim with the threat of publication of content that he has received before, such as private information or intimate photos, and some physical threats to the victim family can lead to a more serious situation than it seemed initially. Assuming that being online is increasingly necessary whether for employment tasks or academic works, is simply not possible to turn off your computer to stop receiving these attacks.

Some studies present a distinction between cyberbullying and cyberaggression. Cyberaggression is defined as aggressive online behaviour that uses digital media in a way that is intended to cause harm to another person. Cyberbullying is one form of cyberaggression that is more restrictively defined as an act of aggression online with an imbalance of power between the individuals involved and repetition of the aggression [11].

3 Related Work

Cyberbullying is a serious social problem especially among adolescents, and it is defined as the use of technology to deliberately or repeatedly attack others. With the emergence of online social networks, this phenomenon has become more prevalent.

Huang et al. [12] used a Twitter corpus to identify social and textual features to create a composite model to automatically detect cyberbullying. They built graphs related to the social network and derived a set of features, to see the context of “me”, “my friends” and the relationship between them, setting weights to the edges to represent interactions between users. The authors indicate that victims of cyberbullying may have a significantly lower self-esteem compared to others and are likely to be more active in networks. They also take a look at the popularity and activity of the users and the number of publications between them. This approach implied the verification of the density of bad words (“asshole”, “bitch”) and hieroglyphs (“5hit”, “@ss”), capital letters rate, the number of exclamation and questions marks, the number of emojis, also checking the part-of-speech (POS) tags and trying to detect text like “you are” or “yourself”.

The study *Modeling the Detection of Textual Cyberbullying* [2] focuses on analysing a corpus of comments from Youtube videos linked to sensitive topics such as race and culture, sexuality, intelligence, or physical appearance. They removed stop words, the non-important sequence of characters (for example, the last character repetitions in “lollll”), and the links for users (e.g., @username). For text classification, they do two experiments: training binary classifiers to check if an instance can be classified for more than one sensible topic; using multi-class classifiers to classify an instance of a set of sensitive topics. They concluded that binary classifiers work best on this problem. The tools used were the Naïve Bayes and Support Vector Machines as classifiers, J48 and JRip as learning methods. In the end, they report that the most difficult sentences to detect are those that contain sarcasm or irony because they do not usually contain the negative words that we are looking for to identify the problem.

Chatzakou et al. [10] characterize a Twitter user by its publications across the time. They classify a user in one of this four categories: aggressive, bullying, spammer or normal. They emphasise the importance to collect some of his profile information like the account age, number of followers and tweet history, especially to check if his kind of speech is constant or if it changed across the time.

Zhao et al. [13] present a mechanism for automatic detection of cyberbullying in social networks through a set of defined bullying features. First, they define a list of bad words based on searches for insulting seeds. Then, based on word embeddings, they extend these linguistic resources to define bullying features, setting different weights to each feature based on the similarity between the word embeddings and concatenating

them with bag-of-words features and latent semantic features to form a vector representation using word2vec. The insulting seeds list contains 350 words that indicate insult or negative emotions (“nigga”, “bitch”, “fuck”, etc.). Using word embeddings, they verify the similarity between words (“beef” and “pork”) through weights assigned to each one of them. When the final representation of each Internet message is obtained, the linear classifier SVM is used to detect the existence of cyberbullying.

The use of images to aid in cyberbullying detection is a complementary approach, given that by posting a photo, an individual may receive insults and provocations from people linked to him on social networks. Also, the publication of intimate photos by others to lead the person on the image to be humiliated in public is one of the concerns to consider in this type of study. Lightbody et al. [3] assure that combining sentiment analysis with image processing techniques is considered an appropriate platform for categorization of textual and visual connotations of content. With this, the authors intend to show that it is not only through text that the attack can be made, since the offensive text can be presented as an image, or else, the offence can be tried by editing a photo. They consider that the most relevant pictures will be those that may contain nudity, evidence of editing and analysis of text within the image. The existence of text related to the image helps to determine the risk of content negativity and the associated category.

Other work [14] intended to analyse the audio and text from Vine platform, as the popular social media sites are becoming increasingly visual, and the use of audio-based interfaces for interacting with both devices and other human beings is constantly growing. The authors believe that cyberbullying grows bigger and meaner with photos and videos, so they decided to collect features from the text, images and sounds of a Vine post such as number of words, sentiment, loudness or presence of drugs. The cognitive aspect of that analysis was related to the identified facial expressions in the images and to the evaluation of the semantic of comments and its similarity to previously identified cyberbullying situations. After detecting a positive case of bullying the objective is to send feedback to some stakeholders like parents or authorities, and trying to hide that from the general public, as they say that identifying individuals as both victims and bullies can have negative consequences, as the victims may be targeted for further bullying and identified bullies may face administrative or legal action.

As Instagram is one of the bigger photo-share social network, Housseinmardi et al. [11] did a manual labelling work for some of this application publications. They refer that cyberbullying on Instagram can happen in different ways, including posting a humiliating image of someone else by perhaps editing the image, posting mean or hateful comments, aggressive captions or hashtags, or creating fake profiles pretending to be someone else. They also consider the common psychological concerns in the available text and in the image content, so they can conclude that it is more probably to be facing a bullying situation when concerns like death or religion are found. Zhong et al. [15] want to look for some specific characteristics in Instagram bullied photos, such as skin tone or outdoor colours, for example, to check the possibility of racism or to find if photos at the beach are more susceptible.

4 Proposed Solution

Due to the increase of Internet usage, with special attention to the usage growth of online social platforms, we believe that action is needed regarding social aspects. More specifically, we want to address the bullying problem on online social platforms. The goal is to build a system to identify situations of cyberbullying autonomously and to adapt to new scenarios that may arise, learning over time and thereby increasing its effectiveness in each prediction for each analysed interaction in an online social network.

Online social networks are commonly used by young people, giving them the possibility to connect and interact with their friends. However, some of their connections can be more unfriendly and may start to execute some practices of bullying to attack someone when they feel anger, envy or jealous in relation to another individual. Knowing that most of the insulting content provided will be found in the form of text, this should be the focus of analysis to classify a situation as bullying or as not bullying.

In a previous work [16] we presented a model for a virtual social sensor that, by capturing the vast amount of public data available in online social platforms, analyses the behaviour of users in social networks. To achieve that, the virtual social sensor contains a set of modules, namely: natural language processing; emotion and sentiment detection; analysis of network interactions; and personality profiling based on the Big Five model [17]. Despite the original purpose of the sensor (being applied in smart environments and organisational applications contexts) with this work we want to endow the virtual social sensor with capabilities to respond to online social hazards.

Despite the virtual social sensor being equipped with a text analysis module, it is necessary to improve it to a point that, based on the similarity between the characteristics of phrases and words considered as offensive, indicates the probability of a new text belonging to each of the two classes mentioned (bullying or not bullying).

After researching about related works that were presented, we could choose to follow the approach to convert the textual content into word embeddings, which will represent it in the form of numerical vectors. That text, after being represented in the cartesian space, will allow us to find a measure of similarity between the different language resources through a Support Vector Machine that can output the final textual classification. However, performing only textual analysis to recognize a bullying situation will not be enough by itself to help reduce these situations in the future, even if the recidivism in practice of these acts can cause the social network to block or delete the aggressor account.

Since the sensor allows us to analyse the interactions in the network, it would be easy to verify what is the kind of people that a user usually interacts more frequently and in what way. For example, if an individual only tends to comment or exchange content with people of the same gender, and if it normally interacts in a way that provokes those people, the online social network may choose to avoid presenting people with a similar profile and personality, as well as not suggesting new connections to people who present these characteristics, in order to avoid scenarios that can lead to new situations of bullying. This may give us more information about the profile of people who are normally involved in a bullying situation, whether it is a victim or an

aggressor, and in the same way that the sensor currently does to present content to the user based on their preferences, there could be an inverse process that would aim not to connect some kind people and some types of publications that could lead to a new bullying situation

There are other ways of practising bullying than just by text. One individual may share a picture of another to try to provoke embarrassment or to make him feel bad. This picture may even be accompanied by a description that further increases the level of aggressiveness of the situation, so another module can be coupled to the sensor to work in these tasks. Having the ability to look for the presence of people in the photos or videos that are usually shared over the different social networks may be an indicator that something is not right. If an image of another person is shared together with an offensive description, it is very likely that we are dealing with a case where someone is trying to threaten or denigrate the person in the picture.

To do this, we need to build a new module that can search for human presence in the images and, after finding them, step into a facial recognition task to know if the person in the image is the same one that shared it. In the case of being a different person who is present in this multimedia content, and the text that is placed in the description fits the patterns of a text characteristic of a bullying situation, the system should indicate this situation as such. This type of implementation will be simplified by the fact that there is several machine learning software available on the Internet that can perform this analysis with a high percentage of success, running its tasks in very short times, and can store the data of previous classifications to constantly improve its performance.

Often these images may contain text written on it due to graphical editions, which is a very common practice in social networks where you just can publish some update if you upload a photo, as the case of Instagram. That text could be the way the bully found to attack its target, even if the subject is not present in that photo. Additionally, some other components may be added to the photo, to make it look different from the reality, trying to embarrass someone. As result, we can also adapt this image analysis module to try to perform Optical Character Recognition tasks in order to convert the text in the image to plain text to be analyzed the same way we want to do with content description and comments.

One capability of the virtual social sensor is being able to identify traits of personality, further leading to the identification and clustering of profiles (based on personality). Since the personality is directly correlated to the behaviour expressed by someone, theoretically, it is possible to identify a bully or a potential situation of bullying before they are expressed. By being able to predict that someone can initiate a bullying situation, it is possible to engage with prevention measures before the hazard occurs, by monitoring users identified with potential bullies, or situations (like public communications) that can lead to a bullying situation.

In this type of cognitive analysis, we need to focus on the type of written text and its frequency, in the categories of posts the user normally interacts with and in the interests he or she may have and can be identified, for example, by the likes on a determined marketing or sports club page. With that, we will easily identify a pattern in user preferences and actions, so by his historical activity, it would be easier to predict if someone can be involved in a bullying scenario.

This solution can be directly implemented on the online social networks platforms to improve their services and stand out as good applications for young people share their interests.

5 Conclusion and Future Work

The constant growth of Internet usage and the services it provides is a result of the increasing commitment of thousands of millions of individuals across the globe. The new realities resulting from online platforms, specifically online social platforms, allow people to socialize and experience different realities, with the expense of some risks. Among those risks, we can find bullying, or cyberbullying, which can be prominent among children and can lead to more serious problems in their lives and having difficulties to be integrated into society.

The ability to identify and monitor these types of situations and identify the profiles of people who are most involved in bullying situations have been explored either by textual analysis or even by image recognition techniques. We believe that current approaches to this topic can be further improved by including a personality characteristic to the analysis.

Motivated by a previous model for a virtual social sensor, we can enhance the text analysis capabilities by including a module for detecting abusive sentences that can be classified as bullying. For the visual aspect, the development of a module capable of image recognition can work alongside the textual content present to enhance the prediction capabilities of possible social hazard situations. This prediction capability is further enhanced by the existing personality profiling module that, in combination with the described modules, can help the identification of new bullies by monitoring behaviour and analysing personality traits similarities with the profiles of known bullies.

As result, our next step is to develop the mentioned modules and work on a model solution for the identification, and categorization, of personality characteristics of bullies, or potential bullies. Since video content is also highly present on online social platforms, we will take into consideration the development and implementation of a sound/video recognition module that can help to identify physical and/or verbal situations of aggression. Implicitly, all this development is associated with a continuous process of data collection to be able to train each module.

References

1. Evans, C., Smokowski, P.: Theoretical explanations for bullying in school: how ecological processes propagate perpetration and victimization. University of Kansas, University of North Carolina, USA (2016)
2. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. MIT Media Lab, Massachusetts Institute of Technology, Cambridge, USA (2011)

3. Lightbody, G., Bond, R., Mulvenna, M., Bi, Y., Mulligan, M.: Investigation into the automated detection of image based cyberbullying on social media platforms. School of Computing and Mathematics, University of Ulster, Northern Ireland. Carnbane Business Centre, Newry, Northern Ireland (2014)
4. Mascheroni, G., Cuman, A.: Net Children Go Mobile: Final Report. Educatt, Milano (2014)
5. McClowry, R., Miller, M., Mills, G.: Theoretical explanations for bullying in school: what family physicians can do to combat bullying. Department of Family and Community Medicine, Thomas Jefferson University, Philadelphia, USA (2017)
6. Hee, C., et al.: Automatic detection of cyberbullying in social media text. Ghent University, University of Antwerp, Belgium (2018)
7. Hardy, R., Norgaard, J.: Reputation in the internet black market: an empirical and theoretical analysis of the Deep Web. *J. Inst. Econ.* (2017). George Mason University, Virginia, USA
8. Patchin, J., Hinduja, S.: Digital self-harm among adolescents. University of Wisconsin-Eau Claire, Eau Claire, Wisconsin, USA. Florida Atlantic University, Jupiter, Florida, USA (2017)
9. The Next Web - "How Dangerous is Cyberbullying?". www.thenextweb-com/contributors/2017/10/04/how-dangerous-is-cyberbullying. Accessed 31 Oct 2017
10. Chatzakou, D., Kourtellis, N., Blackburn, J., Cristofaro, E., Strighini, G., Vakali, A.: Mean birds: detection aggression and bullying on Twitter. In: *Proceedings of the 2017 ACM on Web Science Conference*, pp. 13–22. ACM (2017)
11. Hosseinmardi, H., Mattson, S., Rafiq, R., Han, R., Lv, Q., Mishra, S.: Analyzing labeled cyberbullying incidents on the Instagram social networks. University of Colorado Boulder, Boulder, USA (2015)
12. Huang, K., Singh, V., Atrey, P.: *Cyber Bullying Detection Using Social and Textual Analysis*. ACM, New York (2014)
13. Zhao, R., Zhou, A., Mao, K.: Automatic detection of cyberbullying on social networks based on bullying features. School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (2016)
14. Soni, S., Singh, V.: *See No Evil, Hear No Evil: Audio-Visual-Textual Cyberbullying Detection*. Rutgers University, New Brunswick (2018)
15. Zhong, H., et al.: Content-driven detection of cyberbullying on the Instagram social network. In: *IJCAI*, pp. 3952–3958 (2016)
16. Barbosa, R., Santos, R.: *Online social networks as sensors in smart environments*. CIICESI, ESTGF, IPP School of Technology and Management of Felgueiras, Felgueiras, Portugal (2016)
17. Ghavami, S., Asadpour, M., Mahdavi, M.: Facebook user's like behavior can reveal personality. In: *2015 7th Conference on Information and Technology (IKT), Urmia*, pp. 1–3 (2015)