Artificial intelligence tools in the ongoing fight against bullying and cyberbullying: a multidisciplinary approach

Giulia Orrù[1,*], Vincenzo Gattulli[2], Guido Colaiacovo[4], Stefano Marrone[3], Giovanni Puglisi[1], Lucia Sarcinella[2], Grazia Terrone[5], Donatella Curtotti[4], Donato Impedovo[2], Gian Luca Marcialis[1] and Carlo Sansone[3]

[1] *University of Cagliari, piazza d'Armi, 09123, Cagliari, Italy*
[2] *University of Bari, Via Edorado Orabona 4, 70121, Bari, Italy*
[3] *University of Naples Federico II, Via Claudio 21, 80125 Naples, Italy*
[4] *University of Foggia, Via Antonio Gramsci 89, 71122 Foggia, Italy*
[5] *Tor Vergata University, Via Columbia 1, 00133 Roma, Italy*

Abstract

The problem of bullying and cyberbullying is growing and is a phenomenon that negatively influences our society: for this reason, it requires advanced solutions aimed at prevention. Starting from a previous project called BullyBuster which paved the way for the application of artificial intelligence (AI) in this area, exploiting interdisciplinary skills to develop algorithms capable of identifying and mitigating bullying behaviours, in this paper, we present the follow-up, called BullyBuster 2 (BB2). BB2 aims to broaden the spectrum of intervention to include adult populations, reflecting the universal nature of bullying across all age demographics. Furthermore, it updates and improves the original project's methodologies with new psychological insights and broader AI applications, ensuring a more inclusive anti-bullying strategy. Furthermore, BB2 aims to form a large research consortium to set a new standard for interdisciplinary collaborations in preventive and response strategies against such social challenges.

Keywords

bullying, detection, cyberbullying, artificial intelligence, interdisciplinarity

1. Introduction

The worrying rise in bullying and cyberbullying in recent years has brought attention to the urgent need for efficient prevention measures and solutions [1]. In fact, such violent behaviours not only compromise individual feelings of security and dignity, but they also have significant psychological effects that may reverberate across society. Bullying has changed throughout time and is now prevalent in a variety of adult demographics and settings, which has made it clear that creative and flexible solutions are required.

In this context, the "BullyBuster - A framework for bullying and cyberbullying action detection by computer vision and artificial intelligence methods and algorithms" project [2](BB1 in the following) was proposed: funded under the tender relating to Projects of Relevant National Interest (PRIN) in 2017 and recently concluded, it marked a pioneering effort by interdisciplinary teams from four universities in Southern Italy to tackle this threat through cutting-edge artificial intelligence tools [3]. This initiative sought to combine psychological insights with technological advances to create tools capable of identifying bullying in both physical and virtual contexts and for this reason it has been included in the Global Top 100 list of AI projects addressing the 17 United Nations Strategic Development Goals by the International Research Center for Artificial Intelligence under the auspices of UNESCO[1].

This paper introduces "BullyBuster 2 – the ongoing fight against bullying and cyberbullying with the help of artificial intelligence for the human wellbeing" project (BB2, hereinafter), the follow-up to the BullyBuster project. Funded by the European Union - NextGenerationEU and within the PRIN 2022 PNRR, it again involves four multidisciplinary research groups belonging to four universities in Southern Italy (University of Bari Aldo Moro, University of Cagliari, University of Foggia, University of Naples Federico II) with the aim of combining artificial intelligence, technology, law and psychology skills to develop a comprehensive framework.

BB2's goal is to advance bullying and cyberbullying detection capabilities by incorporating advanced models that take into account a broader range of victim profiles and refining its detection algorithms to accommodate new forms of interpersonal aggression and IT misconduct. BB2 also aims to broaden the objectives of the previous

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

*Corresponding author.
giulia.orru@unica.it (G. Orrù)

[1] https://ircai.org/top100/entry/bullybuster-a-framework-for-bullying-and-cyberbullying-action-detection-by-computer-vision-and-artificial-intelligence-methods-and-algorithms/

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Figure 1: BullyBuster2 project logo.

BB1 to include adults, reflecting the broader spectrum of bullying victims. This initiative is driven by recent social challenges and legislative demands, recognizing that bullying is a pervasive threat not limited to young people but a general public health issue.

A further objective of the project is the creation of an interdisciplinary reference laboratory to combat the phenomenon of bullying and cyberbullying. This laboratory, in which expert technologists, jurists, psychologists, sociologists and economists will participate, will contribute to greater dissemination and expansion of the research community in the field, including stakeholders such as institutions and companies and subjects potentially interested in the engineering of anti-bullying tools.

2. The BullyBuster framework

In this section, we explore the key elements of BullyBuster's architectural framework (Fig. 2), which includes the tools created in BB1 as well as the enhancements planned for BB2. The architecture of the system is structured into five primary modules: (1) a manipulated multimedia content detection module, to combat the spread of malicious and fake multimedia content, as deepfakes; (2) a physical violence detection module, to monitor and identify potential bullying incidents in the physical domain; (3) a stress detection module, able to assesses the emotional state of individuals by analyzing the timing and rhythm of keystrokes and other behavioral biometrics; (4) a verbal abuse detection module, focused on the analysis of textual communications to detect verbal signs of bullying or cyberbullying; (5) a internet addiction detection module, to detect psychological malaise by automatic behavioral analyses. When combined, these components provide a complete system that is suited to address the complex problem of bullying in both physical and digital scenarios. In order to address the critical issues of privacy and data protection, this solutions have been examined from a legal and juridical perspective. Below, we describe the individual modules of the BullyBuster framework.

2.1. Legal, juridical and psychological aspects

Bullying and cyberbullying activities, characterised by aggressive and arrogant attitudes, repeated over time and perpetrated to the harm of children, adolescents, and even adults, have become a critical emergency [4]. Despite legislative efforts [5], research into protecting individuals and their well-being prior to the occurrence of an incident through the prevention of bullying and cyberbullying behaviours is still in its early stages. In fact, intervening in the preventative phase serves to protect the subject's psychophysical integrity and avert victimisation. To ensure the individual's safety, law enforcement activities focused at punishing the perpetrator after the damaging event has happened are no longer sufficient. However, it is critical to respond before a potentially detrimental or dangerous occurrence affects the victims. The methodological approach that BullyBuster adopts is based on the creation of technological tools to combat the phenomenon in an integrated approach, both teleologically oriented towards the protection of psychological profiles based on the well-being of the subject and attentive to legal aspects, with a view to prevention.

2.2. Multimedia contents artifacts

The term "deepfake" refers to any way of creating fake multimedia material representing one or more individuals [6]. The modification may specifically target the individual's identity, speech, or expression. Although this technology is of considerable appeal to some sectors of society, such as the film and video game industries, the harmful applications are vast and particularly concerning. In fact, deepfakes may position a person in non-real humiliating or compromising situations, such as scenes of sex or violence, or force her/him to say something she/he never said. Deepfake can thus be used to insult, stalk, propagate revenge pornography, and abuse or cyberbully someone. In BB1, we exploited the complementarity of several individual video deepfake detectors with appropriate fusion rules to increase the generalization ability of modern deepfake detection systems [7]. The goal of BB2 is to extend the deepfake video detection capabilities and add voice cloning detection to make multimedia content analysis complete.

2.3. Anomaly detection of events

The battle against bullying may be greatly aided by having a visual understanding of human behaviour through video analysis. Behavioural "indicators," whether in the form of victimisation or bullying, might include physical aggression, seclusion, or other physical patterns like encircling. In BB1, systems have been created that can

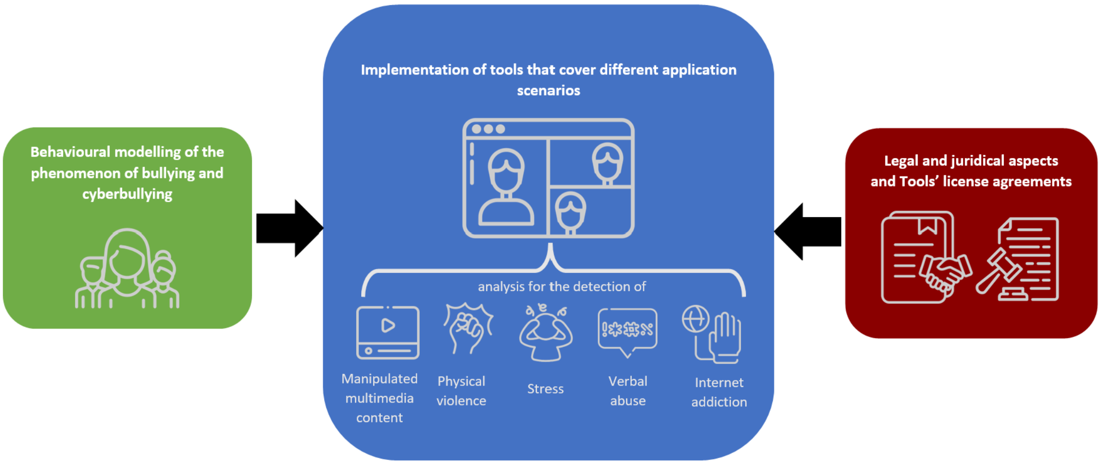
Behavioural modelling of the phenomenon of bullying and cyberbullying
Implementation of tools that cover different application scenarios
analysis for the detection of
Manipulated multimedia content
Physical violence
Stress
Verbal abuse
Internet addiction
Legal and juridical aspects and Tools' license agreements

Figure 2: The BullyBuster 2 project framework.

use one or more surveillance cameras to monitor groups of people who are not individually identifiable [8]. The data is then processed to report "anomalous" events like violent or panicked episodes based on behavioural models that have been suitably codified. In BB2, the crowd analytics model will specialize in bullying contexts by generating synthetic data: obtaining contextualised training data proved to be a challenge during the creation of the crowd analyzer in BB1, made worse by the start of the COVID-19 epidemic, which made it impossible to mimic actual bullying scenarios. In addition, a substantial amount of video is needed for deep-learning models. In BB2, we aim to generate a sufficient amount of data and to include the typical behaviours of bullies and victims, highlighted in the psychological analysis, through the synthetic creation of this data.

2.4. Verbale abuse models

Cyberbullying is a widespread phenomenon involving the use of digital technologies to perpetrate aggressive behavior, threats, or harassment toward other individuals. This bullying occurs primarily through online platforms, such as social media, instant messaging, and forums. It can take many forms, including name-calling, defamation, social exclusion, and disclosure of private information. Cyberbullying can have serious emotional, psychological, and social consequences on victims and poses a significant challenge for parents, educators, and mental health professionals in ensuring a safe and respectful online environment for all users. One of the most predominant vectors is textual language [9].

In the BB1 project, UNIBA focused on identifying verbal aggression, a crucial aspect of cyberbullying that primarily involves texts and comments on online platforms [10]. The goal was to develop a system that automatically identified aggressive behaviour in texts, especially in Italian-language comments. To do this, the language patterns used in aggressive comments were analyzed, noting the frequent use of vulgar and negative language, insults, and offensive words. In addition, many of these comments began with "no/not," which is used to contradict or deny a statement. Based on these observations, several features were developed for feature engineering, including the number of negative words, the number of "no/not," use of capitalisation, positive/negative comment weight calculated with WordNet and SentiWordNet, use of the second person, presence of threats and bullying terms [10]. These features were used to create several dictionaries of terms and a dataset of aggressive and non-aggressive comments. Several shallow learning models were tested to evaluate the effectiveness of these features in detecting verbal aggression [10].

In the context of the BB2 project, this phenomenon can also be analysed by considering different age groups, especially the preponderance of adults. Adults may receive significant criticism, fueling a cycle of online negativity and conflict that can also be reflected in real life. BB2 aims to address this problem by implementing and improving existing text models, with particular attention to the users' ages. While BB1 primarily addressed adolescents, BB2 includes an analysis of adult language, recognizing that verbal aggression and communication dynamics are not limited exclusively to the younger generation. BB2's approach is based on analyzing textual dynamics found in social media comments, examining how different users interact and communicate. This approach allows for the identification of common linguistic and behavioral patterns. The categories used are similarly taken by assigning the label of cyberaggression or non-cyberaggression. Through this in-depth analysis, BB2 aims to provide users, social platforms, and mental health professionals with a better understanding of how verbal aggression occurs online and the factors that influence it. This can enable the development of targeted interventions to prevent and address cyberbullying, improve online safety, and promote a healthier and more respectful virtual environment for all users.

The text also fits within the framework of the Behavioral Biometrics approach, in which touch-related features, called Touch Dynamics, are extracted along with textual output. This could be a starting point for future work [11].

2.5. Stress detection

Biometrics involves the utilization of body measurements and statistical analysis to extract and quantify human characteristics [12]. Initially utilized for user authentication and identification purposes [13], , this technology is now increasingly utilized across various domains, including entertainment and personalized user experiences [14]. Within the field of biometrics, there exist two distinct categories of approaches: i) Physiological biometrics, which involves the direct measurement of physical attributes such as facial features, fingerprints, iris patterns, retinal structures, and vocal characteristics; ii) Behavioral biometrics, which focuses on capturing specific human behaviors such as handwriting, typing, or speaking.

Among various forms of behavioral biometrics, keystroke dynamics has emerged as a highly efficient and economical technique, easily deployable using readily available hardware. In recent years, it has seen increased utilization for user authentication, analyzing the habitual rhythm patterns exhibited while typing on both physical and virtual keyboards [15].

In the context of combating cyberbullying, we employ keystroke dynamics for user emotion recognition. The aim is to exploit its potential as a cost-effective and widely accessible method for emotion recognition, requiring only a standard keyboard as hardware [16]. Furthermore, since a keystroke recorder can be implemented as hardware or software, with the latter being inconspicuous, individuals using the keyboard must be made aware of the monitoring process. In BB2, keystroke dynamics will be examined in real-world settings, considering both keyboard inputs and tapping habits to characterize a subject based on their mobile device usage behavior. Additionally, this analysis can incorporate touch dynamics on mobile devices.

2.6. Digital devices addiction

Examining data from mobile devices has emerged as a prominent approach for understanding individuals' behavior in today's digital era [17]. However, the abundance of features on these devices, such as web browsing, online gaming, digital photography, GPS navigation, and various social applications, can easily engage users' attention, potentially causing significant distractions from real-world activities. Unfortunately, excessive engagement with these features can lead to various health issues, particularly psychological challenges, such as lack of self-control, withdrawal symptoms, sleep disruptions, social isolation, depression, and difficulties in maintaining focus. Additionally, users may display symptoms of irritability, restlessness, stress, and mood fluctuations. In response to these concerns, the goal of BB2 is to introduce an innovative module aimed at assessing individuals from a psychological standpoint, based on their interactions with different features of their mobile devices.

3. The interdisciplinary laboratory

Based on the results of the completed BB1 project and the ongoing BB2 project, an interlaboratory group named "BullyBuster" is being formed with the goal of attracting membership and engagement from individuals who are particularly interested in solving the issue of bullying and cyberbullying. The interdisciplinary laboratory brings together specialists from various fields to develop strategies for analyzing bullying and cyberbullying behaviours, extending beyond just school-aged individuals. This team includes: (1) experts in information security, computer vision, and artificial intelligence, to create and propose innovative models and methods to be implemented in appropriate demonstration products; (2) psychologists, to understand the behaviour patterns of individual involved in bullying events; (3) sociologists, to contribute by examining the effects of bullying and aggressive behaviours in social media and other environments; (4) juridical and legal professionals, to explore new legislation that can address (cyber)bullying, enabling both preventative and punitive technological solutions; (5) economists and healthcare professionals, to assess the financial and health impacts of (cyber)bullying, whether stemming from the attacks themselves or from the origins of such behaviours. The laboratory aims to create a genuine support system and reference services for preventing and fighting bullying.

4. Conclusions

Bullying and cyberbullying are significant societal issues that have an adverse impact on people and communities.

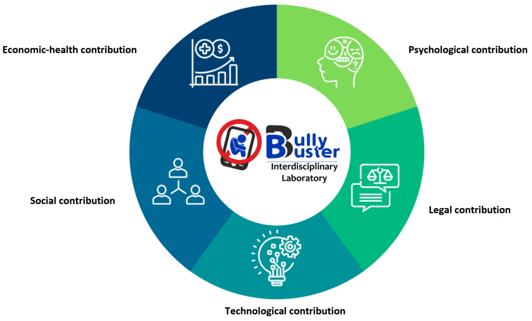
Economic-health contribution
Psychological contribution
BullyBuster
Interdisciplinary Laboratory
Social contribution
Legal contribution
Technological contribution

Figure 3: The BullyBuster inter-laboratory group aims at gathering adhesions and participation of those who are particularly interested in addressing the issue of bullying and cyberbullying according to an interdisciplinary declination.

The BullyBuster (BB) framework, developed during two projects funded under the tender relating to Projects of Relevant National Interest (PRIN), one completed and the other ongoing, combines advanced computer vision, artificial intelligence technology, legal aspects and psychology concepts with a multidisciplinary approach to provide a holistic response to these prevalent problems. In this paper, we present the development objectives of the BullyBuster framework, which include the improvement of tools for the detection of bullying and cyberbullying, the extension of its application to include a target of adults and the creation of an interdisciplinary reference laboratory.

Acknowledgments

This work is supported by the European Union – Next Generation EU under the PRIN 2022 PNRR project “BullyBuster 2 – the ongoing fight against bullying and cyberbullying with the help of artificial intelligence for the human wellbeing” (CUP: P2022K39K8).

References

[1] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, N. Tippett, Cyberbullying: its nature and impact in secondary school pupils, Journal of Child Psychology and Psychiatry 49 (2008) 376–385. doi:https://doi.org/10.1111/j.1469-7610.2007.01846.x.

[2] G. Orrù, A. Galli, V. Gattulli, M. Gravina, M. Micheletto, S. Marrone, W. Nocerino, A. Procaccino, G. Terrone, D. Curtotti, D. Impedovo, G. L. Marcialis, C. Sansone, Development of technologies for the detection of (cyber)bullying actions: The bullybuster project, Information 14 (2023). URL: https://www.mdpi.com/2078-2489/14/8/430. doi:10.3390/info14080430.

[3] G. Orrù, A. Galli, V. Gattulli, M. Gravina, S. Marrone, M. Micheletto, A. Procaccino, W. Nocerino, G. Terrone, D. Curtotti, et al., Leveraging artificial intelligence to fight (cyber) bullying for human well-being: The bullybuster project, in: CEUR WORKSHOP PROCEEDINGS, volume 3486, CEUR-WS Team, Redaktion Sun SITE, 2023, pp. 189–194.

[4] M. Wiertsema, C. Vrijen, R. van der Ploeg, M. Sentse, T. Kretschmer, Bullying perpetration and social status in the peer group: A meta-analysis, Journal of Adolescence 95 (2023) 34–55. doi:10.1002/jad.12109.

[5] M. O. Mantovani, Profili penali del cyberbullismo: la l. 71 del 2017, Indice penale (2018) 475.

[6] M. S. Rana, M. N. Nobi, B. Murali, A. H. Sung, Deepfake detection: A systematic literature review, IEEE access 10 (2022) 25494–25513.

[7] S. Concas, S. M. La Cava, G. Orrù, C. Cuccu, J. Gao, X. Feng, G. L. Marcialis, F. Roli, Analysis of score-level fusion rules for deepfake detection, Applied Sciences 12 (2022). doi:10.3390/app12157365.

[8] G. Orrù, D. Ghiani, M. Pintor, G. L. Marcialis, F. Roli, Detecting anomalies from video-sequences: a novel descriptor, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 4642–4649.

[9] S. Joksimovic, R. S. Baker, J. Ocumpaugh, J. M. L. Andres, I. Tot, E. Y. Wang, S. Dawson, Automated identification of verbally abusive behaviors in online discussions, in: Proceedings of the third workshop on abusive language online, 2019, pp. 36–45.

[10] V. Gattulli, D. Impedovo, G. Pirlo, L. Sarcinella, Cyber aggression and cyberbullying identification on social networks (2022) 644–651. doi:10.5220/0010877600003122.

[11] V. Gattulli, D. Impedovo, G. Pirlo, F. Volpe, Touch events and human activities for continuous authentication via smartphone, Scientific Reports 13 (2023) 10515.

[12] P. S. Teh, A. B. J. Teoh, S. Yue, et al., A survey of keystroke dynamics biometrics, The Scientific World Journal 2013 (2013).

[13] A. Jain, L. Hong, S. Pankanti, Biometric identification, Communications of the ACM 43 (2000) 90–98.

[14] R. L. Mandryk, L. E. Nacke, Biometrics in gaming and entertainment technologies, in: Biometrics in a Data Driven World, Chapman and Hall/CRC, 2016, pp. 215–248.

[15] M. Karnan, M. Akila, N. Krishnaraj, Biometric personal authentication using keystroke dynamics: A review, Applied soft computing 11 (2011) 1565–1573.

[16] S. Marrone, C. Sansone, Identifying users’ emo-

tional states through keystroke dynamics, in: Proceedings of the 3rd International Conference on Deep Learning Theory and Applications - Volume 1: DeLTA,, INSTICC, SciTePress, 2022, pp. 207–214. doi:10.5220/0011367300003277.

[17] F. S. Rahayu, L. E. Nugroho, R. Ferdiana, D. B. Setyohadi, Research trend on the use of it in digital addiction: An investigation using a systematic literature review, Future Internet 12 (2020) 174.