AI and Digital Awareness: An Integrated Approach for the Recognition and Prevention of Cyberbullying Through Visual Content

Alessia Anna Catalano[1,2,*,†], Christian Catalano[3,†], Emanuela Ingusci[1,†] and Danilo Caivano[3,†]

[1] *Department of Human and Social Science, University of Salento, Lecce 73100, Italy*
[2] *Department of Law Studies, University of Salento, Lecce 73100, Italy*
[3] *Department of Computer Science, University of Bari Aldo Moro, Bari 70121, Italy*

Abstract

The CSS – Cyber Social Security project, developed within the extended SERICS partnership (Spoke 3 – Attacks and Defences), proposes an interdisciplinary and technologically advanced approach to preventing cyberbullying and enhancing digital urban security. By integrating Social Sensing paradigms with generative Artificial Intelligence (based on the GPT-4 architecture), the project aims to detect and classify potentially harmful visual content (images and videos) into four categories: racism, body shaming, revenge porn, and happy slapping. The algorithm is trained and validated on empirical data collected through a questionnaire administered to 100 professionals in the educational and psychological sectors. Beyond automated content detection, the model also serves as an educational tool: in school settings, a simplified version of the activity assesses students' digital awareness and fosters critical reflection. The project thus bridges digital security and civic education, offering a participatory and integrated model in which AI becomes not only a technical tool but an epistemological ally for shaping ethical and aware digital citizens.

Keywords

Cyber Social Security, Cyberbullying prevention, Social Sensing, Artificial Intelligence, Digital Education

1. Introduction and Background

The phenomenon of cyberbullying, along with emerging issues related to digital urban security, calls for innovative, integrated, and interdisciplinary approaches. The CSS - Cyber Social Security project , developed within the SERICS (Spoke 3 - Attacks and Defences) extended partnership, aligns with this need. It proposes a model based on integrating the Social Sensing paradigm with the potential of Artificial Intelligence (AI) to analyze and prevent digital social risks. The project is structured across several levels: it collects and analyzes data from social networks, urban sensors, and messaging platforms to map digital risk behaviors[1]. Currently, the project is in the development phase of an artificial intelligence algorithm designed to recognize and prioritize visual content (images and videos) potentially linked to cyberbullying incidents. Once the AI algorithm is fully developed, the collected data will be analyzed and compared with the results obtained from questionnaires, which have been completed by approximately 140 professionals active in cyberbullying prevention. Figure 1 illustrates the workflow of the discussed project.

Bullying and cyberbullying are psychologically and socially significant forms of aggression, particularly among young people. While traditional bullying occurs in physical contexts and is characterized by repeated, intentional behavior and power asymmetry, cyberbullying is marked by its indirect nature, anonymity, persistence, and potentially limitless audience[2]. Common forms include flaming (online insults), denigration (harmful or false content), outing and trickery (non-consensual sharing of personal

Joint Proceedings of IS-EUD 2025: 10th International Symposium on End-User Development, 16-18 June 2025, Munich, Germany.

[*] Corresponding author.

[†] These authors contributed equally.

alessiaanna.catalano@unisalento.it (A. A. Catalano); christian.catalano@uniba.it (C. Catalano); emanuela.ingusci@unisalento.it (E. Ingusci); danilo.caivano@uniba.it (D. Caivano)
0009-0007-6272-3924 (A. A. Catalano); 0000-0003-4038-2317 (C. Catalano); 0000-0001-6633-9937 (E. Ingusci); 0000-0001-5719-7447 (D. Caivano)

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

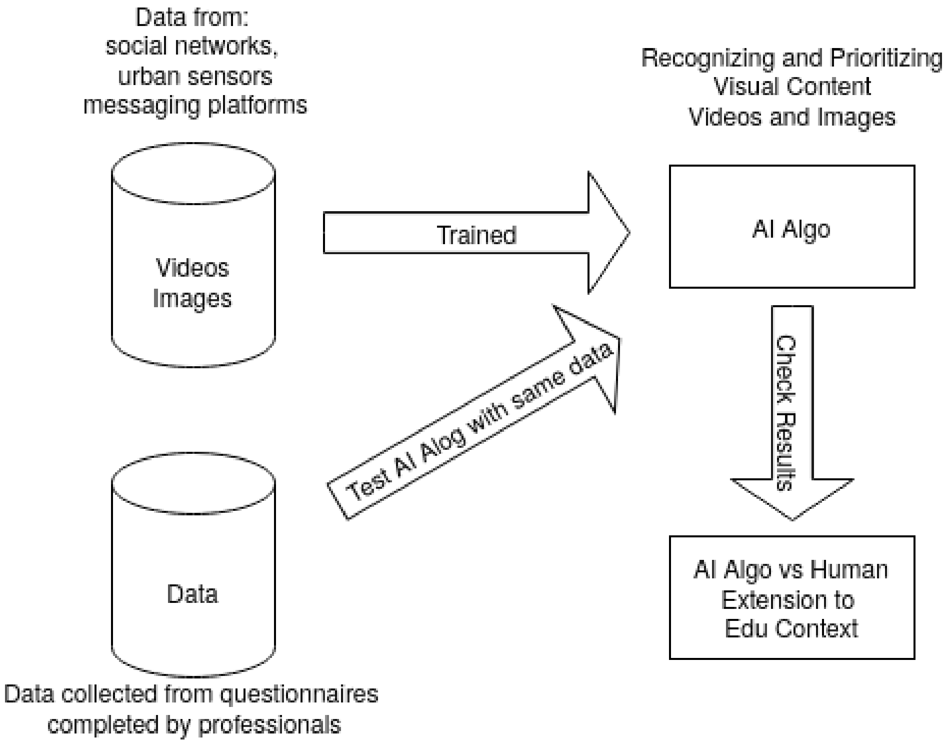
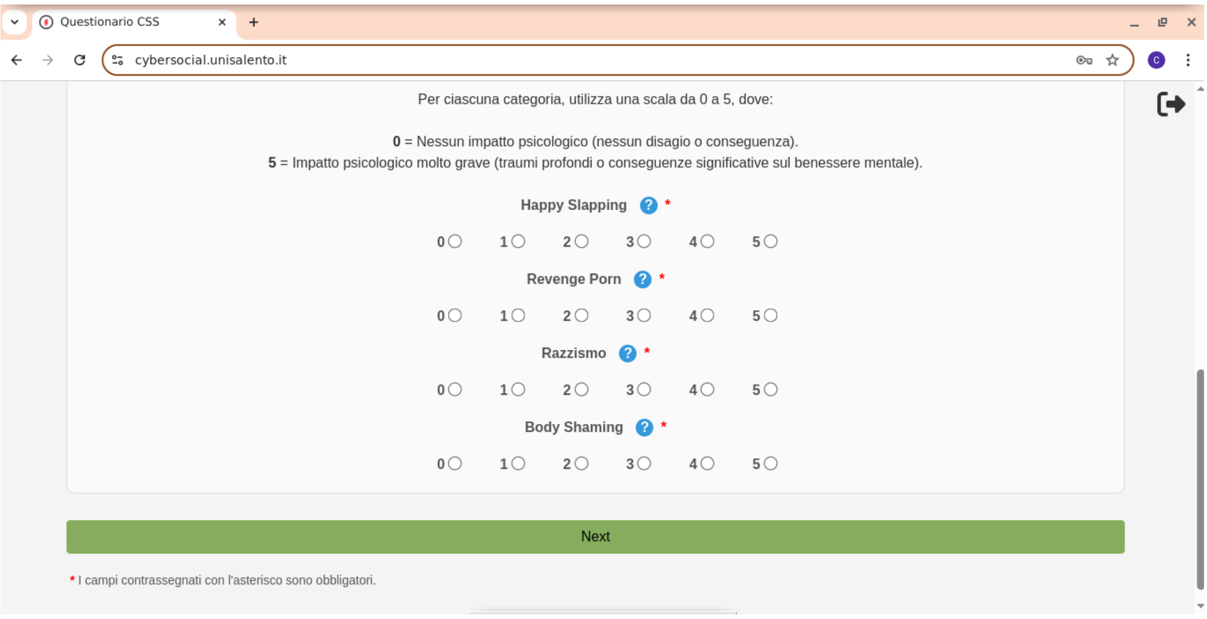


Figure 1: Project workflow

information), exclusion from online groups, and cyberstalking (persistent harassment). Online assaults are pervasive and anonymous, with substantial psychological effects on victims, as highlighted by the European Parliamentary Research Service (2024)[3]. Research (Smith et al., 2008[4]; Nocentini et al., 2010[5]) underscores how the digital environment intensifies and broadens these dynamics. Recent studies (Sultan et al., 2023[6]; Stoleriu et al., 2024[7]) indicate that advanced technologies not only facilitate monitoring and identifying harmful behaviors but also enhance our understanding of their patterns and motivations. Behavioral and motivational taxonomies enable precise classification of aggressors' actions, which is crucial for effective interventions. For instance, Gan et al. (2024)[8] propose classifying digital behavior based on observable patterns or the attacker's intent, differentiating between those aiming to provoke emotional reactions and those seeking to damage reputations. These distinctions inform machine learning models, allowing them to assess not only harmful content but also the underlying intent, enabling more targeted and personalized interventions. The digital context plays a critical role in shaping cyberbullying. Platforms such as Instagram, TikTok, Reddit, online games, WhatsApp, and Telegram serve as spaces where aggression manifests in various forms. Factors such as the victim's identity and the anonymity of the attacker influence the nature and impact of these events. According to the Istituto Superiore di Sanità (2022)[9], many Italian adolescents aged 11–15 have experienced or witnessed bullying or cyberbullying. The pandemic has exacerbated this trend, as confirmed by the 2024 Postal Police Report[10], which recorded a 12% increase in cyberbullying-related offenses over the past year, including threats, image-based abuse, and child grooming. More than 700 investigations were launched in 2023—a 15% increase from the previous year—reflecting improved awareness and international cooperation through over 50 dedicated task forces. Against this backdrop, the CSS project's approach—merging technology, empirical research, and educational awareness—emerges as a model for systematically and strategically addressing cyberbullying.

2. Methodology

The project employs an integrated methodology combining empirical data collection, AI, and social sensing to build an automated system that recognizes and classifies visual content associated with cyberbullying. In the initial phase, a questionnaire was administered—through a specially developed web application (interface of the web application shown in the fig. 2) to 140 professionals (psychologists, educators, and social workers) involved in cyberbullying prevention.



Per ciascuna categoria, utilizza una scala da 0 a 5, dove:

0 = Nessun impatto psicologico (nessun disagio o conseguenza).
5 = Impatto psicologico molto grave (traumi profondi o conseguenze significative sul benessere mentale).

Happy Slapping ? *

0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Revenge Porn ? *

0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Razzismo ? *

0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Body Shaming ? *

0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Next

* I campi contrassegnati con l'asterisco sono obbligatori.

Figure 2: Interface of the web application used to administer the questionnaire to 140 professionals involved in cyberbullying prevention

This provided qualitative and quantitative data on the perception and evaluation of different forms of online aggression. The first section asked participants to prioritize four cyberbullying categories based on their presumed psychological impact: racism, body shaming, revenge porn, and happy slapping. These categories were selected for their visual and semantic recognizability by the AI model.

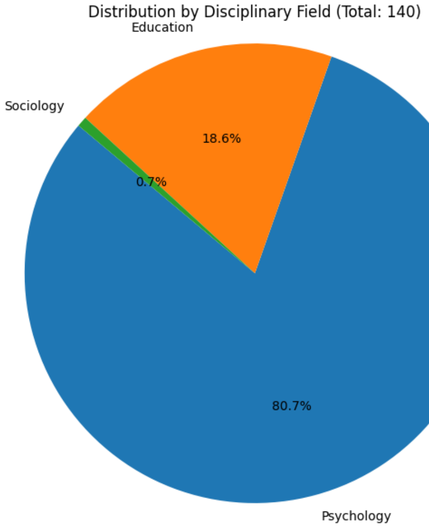


Figure 3: Distribution by disciplinary field.

The sample was predominantly female (81.4%), with male respondents accounting for 17.1% and 1.4%

identifying as otheras.

Regarding professional background, the majority were psychologists (80.7%), followed by educators (18.6%) and a small minority of sociologists (0.7%) as shown in fig.3.

In terms of professional experience, most of the participants (n = 111) had between 0 and 5 years of experience in their respective fields. Whereas, 13 participants stated that they had more than 15 years of experience, 10 between 5 and 10 years and 6 between 10 and 15 years as shown in fig.4.

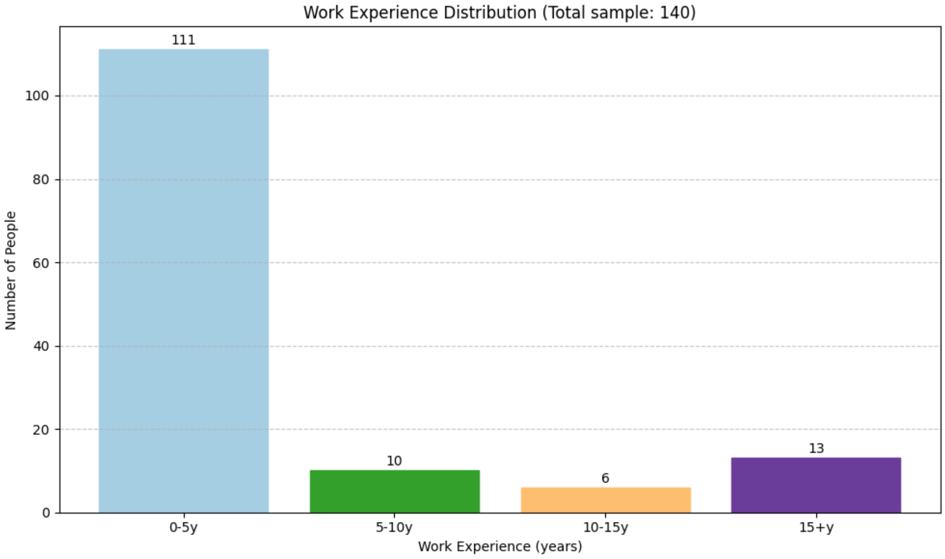


Figure 4: Work experience distribution.

These socio-demographic details provide important context for the interpretation of the evaluative responses and help to adapt the AI system to real-world skills and needs in the field of cyberbullying prevention.

- **Racism**: Discriminatory or violent content based on ethnicity, origin, or skin color, expressed through images, symbols, or actions.
- **Body Shaming**: Offensive or humiliating judgments about physical appearance, often conveyed via memes, videos, or visual comments.
- **Revenge Porn**: Non-consensual distribution of intimate or sexually explicit content aimed at revenge or humiliation.
- **Happy Slapping**: Recording and sharing real physical assaults for online notoriety.

Following prioritization, participants classified a set of images and videos according to the most relevant category. Content was primarily sourced from Telegram channels, known for high-risk material and limited moderation. Each item was pre-analyzed by the research team to ensure clear visual markers aligned with the categories. This dual process — conceptual prioritization and visual classification — enables a systematic analysis of the difference between how professionals in the field (within a specific national context) perceive the phenomenon and how a generative artificial intelligence model (in this case, ChatGPT) does. Expert assessments provide a crucial reference dataset for supervised learning, offering the opportunity to calibrate generative models in the future so that they more closely align with the sensitivity and interpretive criteria adopted by specialized practitioners. The generative artificial intelligence model used in this study is based on the GPT-4 architecture (Generative Pre-trained Transformer 4), developed by OpenAI. These models are trained on vast datasets comprising natural language and, in some versions, multimodal content (including images), allowing them to produce contextually coherent outputs across a wide range of domains. Unlike traditional rule-based systems, generative models do not rely on predefined taxonomies or static knowledge bases; instead, they generate responses dynamically based on statistical patterns learned during pre-training. This architecture makes them particularly effective in tasks involving open-ended reasoning, language generation, and semantic interpretation — but also subject to ambiguity, especially in domains requiring cultural or ethical nuance. Analyzing their behavior in comparison with expert assessments helps reveal not only the strengths of such models in content generation, but also their current limitations in interpretive sensitivity and contextual accuracy. The goal is to integrate human judgment with the computational capabilities of AI, in order to refine the sensitivity, reliability, and accuracy of automatic recognition systems. While currently focused on four main visual categories, the project envisions a progressive expansion of the model to a broader and more nuanced set of behaviors associated with cyberbullying. In parallel, generative artificial intelligence models are being developed not only to improve algorithmic accuracy but also to support awareness-raising and educational initiatives aimed at non-expert audiences. In this context, the versatility of generative AI — particularly its ability to adapt tone, format, and content for different user groups — opens new possibilities for the design of interactive, personalized educational tools. The ultimate goal is to make AI not just a technical tool, but an educational ally capable of promoting awareness, empathy, and digital responsibility in society.

3. Extension to the Educational Context

Several studies are currently being conducted to educate the younger generations on the responsible use of technology, with a particular focus on cyber security[11] and cyber social security[12]. The approach of the CSS project, which integrates expert evaluations with AI-based recognition, holds significant potential for educational applications. Beyond its detection capabilities, the model can be used as a pedagogical tool to assess and enhance students' digital awareness. In school or training settings, a simplified version of the questionnaire can be used. Students analyze and categorize curated visual content according to the four cyberbullying categories. This activity aims to gauge students' sensitivity, risk perception, and interpretative ability regarding harmful digital behavior, rather than training the AI.
This exercise helps:

- Measure digital awareness and identify cognitive, emotional, or cultural vulnerabilities.
- Uncover generational or cultural gaps in interpreting online content, such as the normalization of violence or the underestimation of discrimination.
- Design customized educational paths tailored to students' digital maturity.

Comparing students' classifications with those of professionals and AI results fosters critical and metacognitive reflection. It encourages students to develop ethical awareness, empathy, and civic responsibility in digital spaces. This educational extension illustrates a shift from multidisciplinary to interdisciplinary practice. While project design involves expertise from IT, psychology, pedagogy, law, and communication, the classroom application fuses these fields to create a transformative, learner-centered experience. AI becomes more than a technical tool—it becomes an epistemological partner that actively shapes learning and promotes digital citizenship.

4. Practical Implications

The CSS project has substantial social, cultural, and technological significance, using AI to enhance both digital security and citizenship education. A key innovation lies in focusing on visual content—images and videos—rather than text-based data, addressing a critical gap in current research and tools. This shift responds to the increasing prevalence of visual communication in digital interactions. Memes, GIFs, stories, and short videos often transmit subtle or insidious aggression. Recognizing and classifying this content represents a methodological leap with practical benefits.
The research has dual implications:

- **Technological**: The system provides practical support to platforms, moderators, authorities, and prevention teams. Trained on visual markers and validated by experts, the algorithm reflects how young people communicate online.
- **Educational**: The model serves as a diagnostic and pedagogical tool, enabling customized educational interventions based on students' interpretative abilities. Visual analysis fosters more immersive and relatable learning experiences.

Furthermore, the project frames AI as a bridge between safety and literacy, dissolving disciplinary boundaries to promote systemic and relational thinking. AI, in this context, becomes a partner in education, stimulating critical thinking and civic engagement.

5. Conclusion and Future Directions

The CSS - Cyber Social Security project represents a novel and practical model for addressing cyberbullying and broader digital security concerns. By integrating data, AI, expert input, and education, it exemplifies applied interdisciplinarity and moves beyond traditional siloed approaches. The algorithm, trained on real content and expert-labeled data, is sensitive to both visual indicators and the symbolic-social dimensions of online aggression. Its future use in educational contexts promises to transform AI from a detection tool into a means of fostering awareness, discussion, and growth among students and educators. This research demonstrates both technological efficiency and educational impact. On the technical front, it delivers an advanced tool for analyzing cyberbullying-related visual content. On the educational side, it enables adaptive learning paths grounded in students' digital literacy. Ultimately, AI emerges not merely as a technological asset but as a connector between digital safety and literacy, promoting a more integrated, systemic, and participatory educational model. In this role, AI supports the development of critical, ethical, and engaged digital citizens.

6. Acknowledgments

This work was supported by the following project: National Recovery and Resilience Plan, Mission 4 "Education and Research" – Component 2 "From Research to Business" – Investment 1.3, funded by the European Union – NextGenerationEU – CUP: F53C22000740007 – through the project titled "Cyber Social Security," acronym CSS.

Declaration on Generative AI

Either:
The author(s) have not employed any Generative AI tools.

References

[1] L. Mainetti, C. Ardito, D. Curtotti, T. DI NOIA, A. Corallo, E. DI SCIASCIO, P. Guida, W. Nocerino, et al., Cyber-social security through social sensing: An interdisciplinary approach to cyberbullying and urban security, in: 10th Italian Conference on ICT for Smart Cities And Communities, volume 1, CINI, 2024, pp. 1–2.

[2] S. Çakar-Mengü, M. Mengü, Cyberbullying as a manifestation of violence on social media, Multidisciplinary Perspectives In Educational And Social Sciences Vi 47 (2023).

[3] C. Murphy, Cyberbullying among young people: Laws and policies in selected Member States, Technical Report, European Parliamentary Research Service, 2024. URL: https://www.europarl.europa.eu/RegData/etudes/BRIE/2024/762331/EPRS_BRI(2024)762331_EN.pdf.

[4] 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.

[5] A. Nocentini, J. Calmaestra, A. Schultze-Krumbholz, H. Scheithauer, R. Ortega, E. Menesini, Cyberbullying: Labels, behaviours and definition in three european countries, Australian Journal of Guidance and Counselling 20 (2010) 129–142.

[6] D. Sultan, B. Omarov, Z. Kozhamkulova, G. Kazbekova, L. Alimzhanova, A. Dautbayeva, Y. Zholdassov, R. Abdrakhmanov, A review of machine learning techniques in cyberbullying detection., Computers, Materials & Continua 74 (2023).

[7] R. Stoleriu, A. Nascu, A. M. Anghel, F. Pop, Bullying detection solution for gifs using a deep learning approach, Information 15 (2024) 446.

[8] M. F. Gan, H. N. Chua, M. B. Jasser, R. T. Wong, Categorization of cyberbullying based on intentional dimension, in: 2024 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), IEEE, 2024, pp. 285–290.

[9] Istituto Superiore di Sanità, Health Behaviour in School-aged Children (HBSC): Rapporto 2022 sul bullismo e il cyberbullismo, ISS, 2022.

[10] Polizia Postale, Report annuale 2024 sulle attività di prevenzione e contrasto dei crimini informatici, Polizia Postale, 2024.

[11] C. Catalano, A. Pagano, A. Piccinno, A. Stamerra, et al., Cartoons to improve cyber security education: Snow white in browser in the middle., in: IS-EUD Workshops, 2023.

[12] V. S. Barletta, D. Caivano, C. Catalano, M. de Gemmis, D. Impedovo, Cyber social security education, in: International Conference on Extended Reality, Springer, 2024, pp. 240–248.