International Journal of Management and Humanities (IJMH)
ISSN: 2394-0913 (Online), Volume-4 Issue-11, July 2020

Cyberbullying Detection Through Machine Learning: Can Technology Help to Prevent Internet Bullying?

Jacopo De Angelis, Giulia Perasso

Abstract: ***Cyberbullying constitutes a threat to adolescents' psychosocial wellbeing that developed alongside technological progress. Detecting online bullying cases is still an issue because most of victims and bystanders do not timely report cyberbullying episodes to adults. Therefore, automatized technologies may play a critical role in detecting cyberbullying through the use of Machine Learning (ML). ML covers a broad range of techniques that enables systems to quickly access and learn from data, and to make decisions about complex problems. This contribution aims at deepening the role of ML in cyberbullying detection and prevention. Specifically, the following issues are addressed: i. identifying the features most frequently considered to develop ML models predicting cyberbullying; ii. identifying the most used ML algorithms and their evaluation methods; iii. understanding the implication of ML for prevention; iv. highlighting the main theoretical and methodological issues of ML algorithms in predicting cyberbullying. To answer these research questions, a systematic review of literature reviews, from a total of n=186 records from online databanks, has been conducted. Ten literature reviews have been elected to analyze and discuss evidence about ML preventative potential against cyberbullying. Most of the models used content-based features to predict cyberbullying. The majority of these features includes words written in social network posts, whereas Support Vector Machine, Naïve Bayes, and Convolutional Neural Networks are the most used alghorithms. Methodological and technical issues have been critically discussed. ML represents an innovative preventative strategy that may optimize and integrate educational programs for adolescents and be the starting point of the development of technology-based automatized detection strategies. Future research is challenged to develop algorithms capable of detecting cyberbullying from several multimedia sources.***

Keywords: Cyberbullying; Machine Learning; Cyberbullying Detection; Systematic-Review; Prevention.

I.INTRODUCTION

In the last decade, cyberbullying has become a severe threat to adolescents' psychosocial wellbeing [1, 2]. Cyberbullying consists of intentional harmful behaviors towards a target-victim, enacted via electronic devices [3]. Nevertheless, cyberbullying is not only a transposition of bullying into cyberspace because of its distinctive characteristics [4]. For example, repetitiveness is a core component of traditional bullying dynamics, whereas it is not essential in cyberspace where posting a content only once can count as multiple times [5,6].

Revised Manuscript Received on July 15, 2020.
* Correspondence Author
Jacopo De Angelis*, Department of Psychology, University of Milano-Bicocca, Milan, Italy. Email: j.deangelis2@campus.unimib.it
Giulia Perasso, Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy. Email: giulia.perasso01@universitadipavia.it

Many theoretical frameworks shaped psycho-social research around cyberbullying, such as: the General Strain Theory (GST), the Routine Activities Theory (RAT), and the Social-Ecological-Theory [7,8,9,10,11]. According with these interpretations, cyberbullying is a multidimensional construct encompassing a broad set of online behaviors [12]. Several taxonomies describe flaming, harassment, denigration, impersonation, outing/trickery, exclusion, cyberstalking, revenge porn as forms of cyberbullying [13, 14, 15, 16]. In such a broad categorization, Menesini and colleagues have traced a line between two main forms of cyberbullying: i. written cyberbullying in the form of verbal offenses; ii. visual cyberbullying, perceived by adolescents as more harmful than written cyberbullying, consisting in non-consensually sharing denigrating videos or pictures [17]. Research also remarked that cyberbullying is a threat to adolescents' psychosocial wellbeing, characterized by symptoms of stress, depression, and anxiety, which are more severe than those observed in traditional bullying [18,19]. Furthermore, several empirical studies have been taking into consideration the link between the types of adolescents' online activities and cyberbullying [20, 21]. In particular, the use of Social Network sites and gaming platforms are the activities associated with a higher risk for cyberbullying [22,23,24]. For these reasons, research is currently stressing the importance of identifying the protective factors against cyberbullying to structure tailored primary, secondary, tertiary prevention programs [25]. Technology could significantly contribute to the development of these preventative strategies by the use of Machine Learning algorithms (ML).

A. Machine Learning and Its Applications

Machine Learning (ML) covers a broad range of techniques that enables systems to quickly access and learn from data, and to make decisions about complex problems. The application of ML has also been growing in many different fields, such as biology, genetics, marketing, medical, and psychological sciences [26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37].
Research clustered ML algorithms into three categories:

- Supervised: The family of supervised ML techniques assumes that labeled continuous or categorical outcomes are available [38]. For instance, in models with categorical outcomes, the model learns patterns from given input data or features (e.g., age, gender, social status etc.) and uses them to classify new observations into a category (diagnostic group, consumer behavior etc.).

Popular classification algorithms include k-nearest neighbor (KNN), Naïve Bayes, decision trees, and random forests.

- Unsupervised: These ML techniques assume that data used to train the model are not labeled or classified. These algorithms are commonly used to describe the data structure when outcomes are not *a-priori* defined. An example of an unsupervised ML algorithm is *K*-means, which is one of the most common clustering algorithms. Given a set of features, *K*-means aims to partition data points into *K*-clusters. The underlying rationale is that each data point in a cluster is more similar to points of its own cluster than to points from other clusters [38].
- Deep learning: these algorithms reproduce the structure and the functions of biological neural networks [39]. These computing systems include deep learning, reinforcement learning, associative rule learning, and convolutional neural networks (CNNs). The common aim of deep learning techniques is to reproduce and implement cognitive, learning, perceptual mechanisms of the human brain into artificial systems, designed to solve several tasks. For example, CNNs' basic structure simulates the functioning of the human visual cortex to perform face recognition tasks and face classification [40, 41].

B. Machine Learning and Cyberbullying

According with the Pew Internet and American Life Project, 80% of adolescent Internet users are social networks site users [42]. The high number of adolescents online has been leading to an increase in adolescents' risk for negative, unhealthy, and dangerous online experiences [43]. Literature describes several interventions to prevent cyberbullying [44]. However, the activation of secondary prevention programs is still an issue because most of victims and bystanders do not report cyberbullying episodes to adults [45]. Therefore, ML algorithms could serve to early and automatically detect cyberbullying, and to foster intervention protocols' timely activation. In the last decade, researchers have tested a variety of ML techniques such as victims' sentiment informed analysis, textual, and semantic analysis, and user features' analysis (e.g., gender) [46, 47, 58, 49]. Plus, these methods allowed to detect a variety of cyberbullying outcomes including binary classifications (e.g., being or not-being involved), role identification in cyberbullying dynamics and the severity of consequences [50, 51, 52]. Despite the opportunities that these algorithms provide in terms of prevention and intervention, social scientists and education institutes seem to overlook the potential contribution of these techniques. Accordingly, this work intends to increase scientific community's awareness about the benefits of ML for the prevention of cyberbullying, by fostering a theoretical and empirical connection between computer science and social sciences.

C. Aims and Research Questions

The present work aims at collecting and discussing research about the use of ML for cyberbullying detection by examining findings from previous reviews. Focusing on reviews enables to highlight the contemporary debate and theoretical views of the topic, addressing the following research questions: i. What are the features most commonly considered to automatically detect cyberbullying?; ii. What are the ML techniques (i.e., algorithms), and how are they evaluated?; iii. What are the implications of ML for prevention?; iv. What are the main issues of ML algorithms to predict cyberbullying?

II.METHOD

The present review implemented a systematic search strategy and selection process [53] (see Figure 1). Records were collected from Scopus (*n*=98), PsycInfo (*n*=88), PsycArticles (*n*=2), finally obtaining *n*=188 records.

Scientific databanks research was conducted in April 2020. A filter of

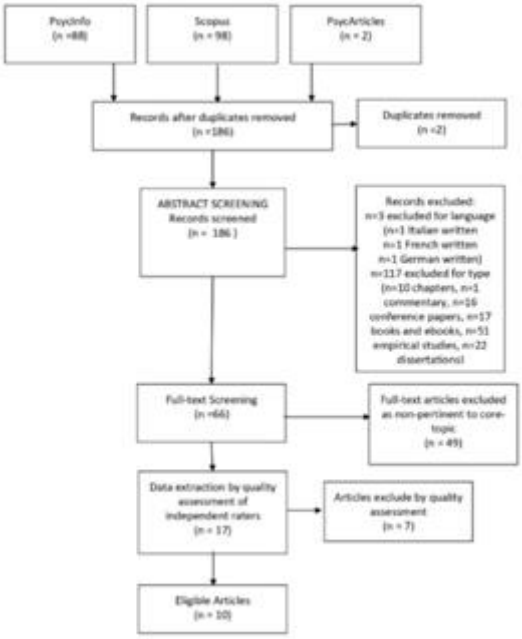

Fig 1. Selection Diagram.

keywords and Boolean operators was used to search across databanks, including the key-words "*cyber*bullying*" (OR online bullying, OR cyber*victimization), AND "*machine learning*" (OR *artificial intelligence*), AND "adolescen*" (OR *teen**), AND "*review*". Two duplicates were removed from the total before abstract screening, obtaining *n*=186 records. From the abstract screening, only English-written records were considered favorably for inclusion, while three sources were excluded (*n*=1 Italian-written, *n*=1 German-written, *n*=1-French-written). Subsequently, a selection based on the type of publication was applied to select only review papers: *n*=10 chapters, *n*=1 commentary, *n*=16 conference papers, *n*=17 books and e-books, *n*=51 empirical studies, *n*=22 dissertations were excluded from the process. The remaining 66 records were full-text screened by applying a topic pertinency criterion based on the paper's core-topic. The criterion allowed to select only records focused on cyberbullying, machine learning, and adolescence. The 17 remaining records were evaluated by two separate raters who applied a quality assessment on data-extraction.

An *ad-hoc* tool was used for quality assessment due to the interdisciplinarity of the sources (for further information see Appendix A). After inter-rater independent data-extraction, *n*=10 review-papers resulted eligible for the present systematic review.

III.RESULTS

The ten selected reviews were extracted by two independent raters, obtaining a substantial inter-rater Pearson's correlation (*r*=.91). As reported in Table I, 60% of the selected reviews were published during 2019, while only *n*=1 review was published in 2018, *n*=1 in 2016, *n*=1 in 2015, and *n*=1 in 2013. Review studies originate from different countries (*n*=2 Malaysia, *n*=1 Israel *n*=1 Saudi Arabia, *n*=1 Switzerland, *n*=1, Turkey, *n*=1 South Africa, *n*=1 Italy, *n*=1 Portugal, *n*=1 India), with a slight preponderance of Malaysian and European sources. As regards the consulted sources, five reviews did not report the exact number, whereas the other five reported the exact number of consulted sources (ranging from a minimum of 3 to a maximum of 22). Finally, all the included literature reviews highlight the importance of ML to prevent and counteract cyberbullying but also point out several issues in the implementation of such techniques.

Table- I: Information of selected records.

Authors	Year	Nation	Databanks (*n*)	Main Findings
Al-Garadi et al. [55]	2019	Saudi Arabia	9	Prediction models can be used to detect and counteract online aggressive behaviors.
Ashikin and Norhalina [70]	2016	Malaysia	Not-Specified	Deep networks techniques and Convolutional Neural Networks (CNN) are efficient in classifying violent videos on multiple internet environment.
Blaya [73]	2019	Switzerland	22	ML techniques are listed among the main areas of prevention against cyberbullying. However, most of these preventative/intervention strategies focus on victims and not aggressors.
Can and Alatas [58]	2019	Turkey	Not-Specified	Social network analysis through ML is applied to 21 online social networks' problems, including cyberbullying detection.
Fire et al. [74]	2019	Israel	Not-Specified	Online social network (OSN) users face many threats for their security. Machine Learning is efficient in predicting spammers and fake profiles.
Mahlangu et al. [60]	2018	South Africa	12	Limited availability of accessible public datasets limits the implementation of ML methods in cyberbullying detection.
Nadali et al. [56]	2013	Malaysia	Not-Specified	Support vector machine (SVM) classifiers are efficient in detecting cyberbullying in chat rooms and forums.

Nocentini et al. [75]	2015	Italy	3	ICT tools are underused in cyberbullying detection.
Rosa et al. [64]	2019	Portugal	6	The lack of coherence in the definition of cyberbullying impacts on detection methods and prevention strategies.
Singh and Kaur [57]	2019	India	Not-Specified	ML methods are used to detect content-based cybercrime and cyberbullying

IV.DISCUSSION

The selection process led to select ten literature reviews addressing the main purposes of the present work. The next paragraphs will critically discuss the following key-points inferred from the selected sources concerning ML applications: i. Main features considered by previous reviews to build predictive models of cyberbullying; ii. Main algorithms and metrics used to evaluate models' performance. iii. Implications of ML for cyberbullying prevention; iv. Main issues will be considered and debated.

A. Main Features

Five of the selected reviews provided an overview of the features that are commonly associated with cyberbullying detection (see Figure 2 for details). In ML models, a feature is an input variable referring to a measurable property of an observed phenomenon (e.g., the content of an image) [54]. The most common selected features in cyberbullying predictive modeling are content-based features [55, 56]. A typical content-based approach consists of an analysis of the valence of the words written in posts on social networks. Text words are extracted from social media and used as input variables (or predictors) to predict cyberbullying. For instance, swear words are a prototypical example of words that substantially predict cyberbullying [57]. Al-Garadi and colleagues (2019) and Singh and Kaur (2019) showed that bag-of-words (i.e., text data analysis where a text is segmented in a "bag" of its words) is the most frequent approach used to analyze texts valence for cyberbullying detection [55, 57]. In this approach, texts are segmented in a "bag" of words, in which texts are transformed into a vectorized word count. In other words, texts are processed as vectors in a way that mathematical operations can be easily performed. These natural language-based approaches also enabled to extract behavioral features from online conversations [56]. These features comprise both observed users' behavioral patterns (e.g., number of questions) and his/her latent intentions (e.g., humiliating, grooming etc.). These behavioral patterns can be inferred by computing specific metrics. For example, *term frequency* is an index of how much frequently a word appears in a document, given by the ratio between the target word count and the total number of words in the text. On the other hand, *the inverse document frequency (IDF)* computes the importance of a word in a group of texts. Specifically, given a total number of documents "A", and the number of documents D in which a word "W" appeared (i.e., "B"), IDF is the logarithm of the ratio between A and B.

The use of these computational techniques has shed light on the posts' themes associated with cyberbullying. According to Can and Atlas (2019), cyberbullying mainly occurs when social media texts include specific conversation themes such as death, appearance, religion, and sexual content [58]. This aspect might be exaggerated by the disinhibition effect provided by the Internet environment, which leads users to express more violently than they would do into face-to-face interactions [43, 59].

One of the selected reviews underlined that most of the literature has focused on detecting cyberbullying from text data rather than images [60]. This imbalance might be due to a limitation of text data compared to images or videos in the stored data availability. This issue is even more relevant given the evidence on the existence of two distinct patterns of cybervictimization (written and visual) [17]. Therefore, special attention has to be paid to how these two distinct forms of cybervictimization are differently conveyed on social media sites. For example, cyberbullying attacks *via* images and videos might mostly be conveyed on Tik-tok and Instagram, since these sites mainly exhibit visual contents. On the other hand, text cyberbullying may be more common on Twitter and Whatsapp, because the majority of contents of these platforms are written. Given that adolescent internet users perceive visual cyberbullying as more dangerous than textual, ML applications seem not to encounter yet the urgency to investigate image and video cybervictimization [17].

Finally, a minority of the studies took into consideration profile-based features such as user profile information (e.g., age and gender) and social media information (e.g., number of likes and followers). As for the former ones, evidence from psychosocial studies stresses the necessity to include personal information such as gender [56]. Previous studies have indeed demonstrated that male and female users act in a different way when involved in cyberbullying dynamics [61, 62]. For example, female users tend to adopt aggressive communication styles (e.g. excluding other users from a group or conspiring against them), whereas men tend to use more threatening words. On the other hand, as pointed out by Can and Atlas (2019), social media information has provided interesting hints for a better understanding of the online user behaviors associated with cyberbullying [58].

They reported that posts with cyberbullying contents received comments more frequently and fewer likes per post than other posts [58]. This evidence might find an explanation by referring to the prototypical roles of cyberbullying involvement. In particular, the Theory of Planned Behavior categorizes three types of cyberbullying bystanders: the one who joins the bully in attacking the victim, the one assuming a neutral behavior, and the one helping the victim [63]. Thus, a possible speculation might be that users commenting on cyberbullying posts, without liking them, are those who are involved in the cyberbullying dynamics without being perpetrators or victims. Precisely, victim defenders represent the phenotype fitting more this description. However, a deepened analysis of the contents of these comments is fully recommended.

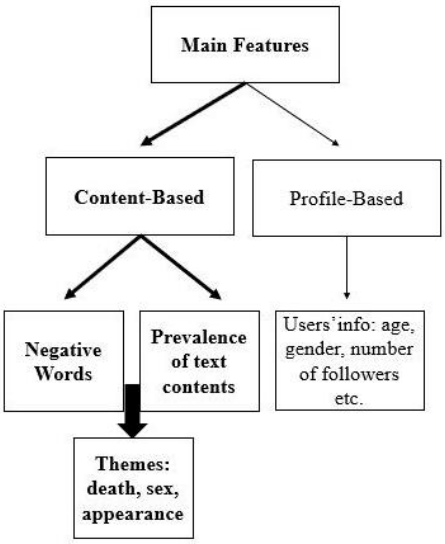

Fig 2. Main features considered by previous studies to build predictive models of cyberbullying.

B. Main Algorithms

Six of the selected reviews reported the Support Vector Machine (SVM) algorithm to be the most used ML algorithm in cyberbullying detection [55, 56, 57, 58, 60, 64]. In this model, data points are separated by a hyperplane fitting the model in a way that two distinct classes are identified. The algorithm searches for the best hyperplane by selecting the two datapoints (i.e., support vectors), which are closest to the hyperplane. The distance between the hyperplane and the two support vectors is named Maximum Margin. Accordingly, the goal of the algorithm is to identify the hyperplane, which maximizes the Maximum Margin, by improving the classification capacity of the algorithm. Instead of a linear hyperplane, non-linear boundaries can be applied in problems in which a linear solution does not fill the data structure (i.e., the so-called "Kernel trick") [65]. Specifically, the following steps have to be followed to perform SVM:

1. A linearly separable hyperplane is sought to separate the values of one class from the values of the other class (i.e., cyberbullying vs no cyberbullying). If more than one hyperplane exists, the one with the highest margin (Maximum Margin) has to be chosen, as it guarantees better accuracies.
2. If such linear solution is not found, data are transformed into a higher dimensional space by exploiting the Kernel trick. Through the Kernel trick, 2D data are mapped into a 3D structure.
3. In order to choose the best boundary (e.g., line in linear problems, or circle in non-linear problems), parameters' optimization has to be performed. In detail, an important parameter has to be tuned in all of the SVM algorithms that is C, the misclassification cost. In fact, if C is exaggeratedly high the model will result at risk of overfitting (variance), whereas if C is low the model will be at risk of underfitting (bias). Accordingly, optimum C has to be found to meet the bias-variance tradeoff.

Three main reasons underlying the popularity of this algorithm in this field can be retrieved: i. cyberbullying outcomes are commonly operationalized as binary outcomes (e.g., bullied or not bullied) as shown by previous research [66, 67, 68]; ii. Rosa et al. quantitively demonstrated that the SVM is the best performing algorithm by supporting its use in cyberbullying detection [64]; iii. the applicability to linear and non-linear data distributions represents a remarkable advantage of this algorithm.

As pointed out by three of the selected reviews, the Naïve Bayes (NB) algorithm is another commonly used algorithm for cyberbullying detection on social media [55, 57, 58]. The NB is a ML algorithm relying on the implementation of Bayes theorem (Figure 3).

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Fig 3. Bayes Theorem.

The model aims to compute the posterior probability of an event A given an event B (P(A|B)), based on prior knowledge (P(A)), on the calculation of the probability of the event B given the event A (likelihood), and on the marginal likelihood (P(B)). In the case of cyberbullying, the posterior probability can be defined as the probability of a cyberbullying event given a set of features (e.g., words), and previous knowledge on the phenomenon (e.g. number of cyberbullying episodes occurring online). It is called naïve since it assumes that features are independent from each other, which is a condition intrinsically unreachable in most of the problems. As regards cyberbullying, the words contained in a social media post are a representative example of features that are unlikely to be independent from each other. As reported by Al-Garadi and colleagues, NB is a high-speed algorithm that is very well suited for text classification problems such as posts content detection [55]. However, the collected evidence suggests that SVM is a more accurate classifier than NB [55]. In line with the "No Free Lunch Theorem", running both algorithms may be the best choice since the success of an algorithm varies as a function of the specific type of problem and the implicated variables [69].

Despite the popularity of SVM algorithms in cyberbullying detection, Nadali et al. and Ashikin and Norhalina showed that deep learning may potentially perform better in this field [56,70]. In particular, Ashikin and Norhalina reported the advantages of applying Convolutional Neural Networks (CNNs) for violent video classification. CNNs are a variant of traditional Artificial Neural Networks designed to identify visual patterns from pixels of an image at a low cost in terms of preprocessing [70]. Specifically, several feature maps are computed to detect the distinctive elements of a picture. The creation of the feature maps is generated iteratively by creating several convolutions. Static frames, motion, and audio features are extracted and integrated, by flattening the feature maps, to predict violent contents in online videos with the use of traditional neural networks. An increase in the use of this technique is expected to reduce the gap between visual and written cybervictimization. Figure 4 shows a schematic depiction of these findings.

C. Main Metrics

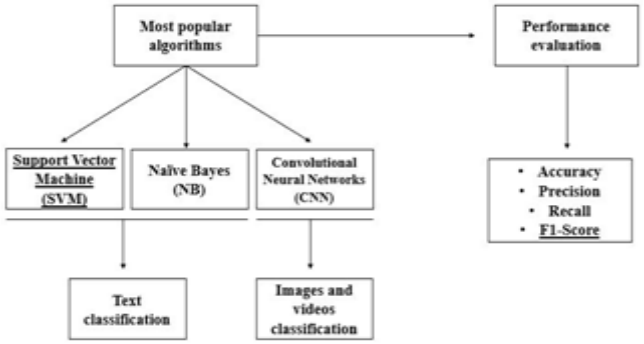

Fig 4. Main algorithms and evaluation metrics used for cyberbullying detection.

Metrics are a set of statistics used to evaluate the performance of ML algorithms. Although several types of metrics exist, their applicability relies on the measurement scale of the outcome variable. For instance, the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) are the most common metrics used for ML regression algorithms, namely when the outcome is continuous. In contrast, accuracy, sensitivity, and specificity are metrics commonly used for classification problems (i.e., models containing categorical outcomes), such as the case of cyberbullying – often operationalized as a categorical variable. Given the classificatory nature of models predicting cyberbullying, evaluation metrics of classification models are the most applied in this field [55, 57, 64]. In detail, three of the selected reviews consistently reported four frequently used metrics, that are accuracy, precision, recall, and F1 score. It is possible to describe these metrics as measures of the classification task performance, based on the concept of true positive (TP; i.e., correct classification of cyberbullying contents) vs. false positive (FP; i.e., non-cyberbullying contents classified as cyberbullying), and true negative (TN; i.e., correct classification of non-cyberbullying contents) vs. false negative (FN; i.e., cyberbullying content classified as non-cyberbullying). Accuracy is the ratio between the sum of TP plus TN observations and the total number of cases. Precision is the ratio between TP observations and the number of total positive predicted values (both FP and TP). Recall (or sensitivity) is the ratio between the TP observations and the sum of TP and FN observations. In other words, precision can be exemplified as "How many contents classified as cyberbullying were truly cyberbullying?". In contrast, recall can be expressed as "How many true cyberbullying contents were correctly classified?". Recall is a measure typically used in medical diagnosis, which defines the sensitivity of a diagnostic tool, that is the ability to correctly detect a specific pathological condition [71]. Finally, the F1 score is the result of the harmonic mean between precision and recall [72]. Accordingly, the F1 score is a more comprehensive metric that includes both recall and precision information. Rosa et al. stated indeed that, despite the relevance of recall as a key-metric in cyberbullying detection, the F1 score is the most balanced way to evaluate cyberbullying classification algorithms [64].

Overall, it is possible to conclude that different evaluation metrics relate to different research and applicative questions. Recall seems indeed more indicated to evaluate the diagnostic or screening reliability of the algorithm, whereas precision may be more appropriate to evaluate its technical performance accuracy. Accordingly, the F1 score could be intended as a cross-cutting global accuracy measure. Figure 4 shows a schematic depiction of the main metrics adopted to evaluate algorithms' performance.

D. ML and Links to Prevention

Five of the selected reviews considered the potential practical use of ML algorithms as well as their link with prevention. According to Blaya, the automatic identification of hateful contents or online bullying behaviors is part of cyberbullying prevention [73]. Identifying online violent acts might indeed enable an earlier implementation of prevention policies and interventions [73]. Plus, these automatic detection systems can substantially contribute to detecting those cyberbullying cases which are more difficult to spot by parents, teachers, clinicians, or social workers. The algorithm, in this sense, may be intended more as a support tool for professionals, teachers, or caregivers, than an automatic detector that only spots and removes hateful contents from social media. The need for integration between computer science and social sciences' interventions is also stressed by Singh and Kaur, who underlined how previous social and psychological literature had neglected the contribution of technology to cyberbullying prevention [57]. In this sense, these algorithms may lead to the development of applications, usable by school staff, able to detect suspicious interactions or contents, and immediately warns school leaders and counselors. Rosa et al. also stressed the importance of using ML algorithms for prevention purposes, by focusing on the need to integrate digital tools (e.g., applications, games, and websites) in social networks to prevent cyberbullying [64]. These systems should be improved at detecting cyberbullying at its earliest stage, by preventing cyberbullies from reaching their target [64].

The authors focused on showing the so-called reflective interface, which is a recent approach addressed at assisting users in making positive decisions based on empathy and prosociality. Fire and colleagues reported a variety of solutions for users to protect themselves from cyberattacks [74]. These solutions include both automatic detection systems based on ML algorithms (e.g., fake profile detection, phishing detection etc.) and good online habits, called operator solutions (e.g., privacy settings, security system, and antivirus updating). Plus, Nocentini and colleagues mentioned an automatic detection system targeted at detecting bullying contents in instant messages (i.e., the MISAAC system) [75, 76]. Specifically, MISAAC uses a traffic light-like system that automatically classifies posts' authors as green, yellow, or red, based on the aggressiveness escalation of contents. MISAAC could be classified as a type of secondary prevention as it detects cyberbullying interactions by preventing them from getting worst [76].

E. Main Issues

The main issues of the application of ML techniques for cyberbullying detection relate to three areas: theoretical, technical, and methodological.

- The most significant theoretical issue concerns the lack of consensus in defining cyberbullying as stressed by two of the selected sources, which reflects the contemporary need to reconceptualize cyberbullying [4, 58, 73]. In literature, it is not clear yet whether repetitiveness of cyberbullying behaviors is a critical component of cyberbullying or not [6]. Accordingly, algorithms do not detect the cumulative frequency with whom a user performs or is the victim of cyberbullying attacks, which raises the need for algorithms predicting the intensity and the severity of cyberbullying. Moreover, persons do not have the same permeability to adverse events, and the same words have different meanings expressed in different contexts. Sharing a standard definition of cyberbullying might prevent biases related to the subjective and contextual perception of the phenomenon, enabling the selection of relevant features (e.g., harmful words) and developing systems capable of encountering the user's needs [73]. Another possible direction may be considering the application of ML to the different types of cyberbullying (i.e., flaming, harassment, denigration, impersonation, outing/trickery, exclusion, cyberstalking, revenge porn, etc.) included in the most relevant taxonomies [13, 14, 15, 16].
- The authors of the selected reviews raised two main technical issues related to online databanks [60, 64]. The higher proportion of text than images data represents the first limitation of the detection systems. Multimedia contents are the primary type of files shared on social networks (e.g., images with caption, video live streaming, and audios) [60]. The lack of these contents depends on the limited availability of datasets containing images data. However, it represents a substantial empirical and theoretical limitation, given the evidence showing that visual cybervictimization is considered more aversive than textual [17]. A second limitation is the lack of information on how datasets were built [64]. Most of the studies do not provide guidelines for the annotators to label the data samples. Likewise, inter-rater reliability and annotators' expertise are often omitted [64]. This issue reflects the absence of a common definition of cyberbullying, as reported before ahead, and of shared and transparent procedures of data annotation [4]. As a consequence, it substantially increases the risk of rater subjectivity biases, that relies on the heterogeneity of cyberbullying definitions and on the criteria used to categorize a text content as cyberbullying [55].
- Methodological issues primarily relate to the external validity problem. External validity is the degree to which the results of a study have an actual impact outside the context of that study [77]. Although ML algorithms enable fast detection of bullying in cyberspace, no evidence has assessed the secondary impact of these systems on users' psychosocial outcomes yet (e.g., mood, anxiety, etc.). Even though many of the selected reviews outlined the link between fast-automatic detection and activation of supportive intervention strategies, it is not clear whether and how these interventions may occur. This problem highlights the necessity of integration between technical and social intervention strategies. Additionally, the impact of these systems on users' real-life should rely on objective and rigorous protocols of assessment. Nocentini and colleagues reported that many Information Communication Technology (ICT) interventions against cyberbullying (e.g., MISAAC) do not provide sufficient information related to statistical effectiveness of the intervention, by only relying on descriptive information such as user's degree of satisfaction [75].

F. Future Directions

Research should aim at developing, training, and testing ML classifiers detecting cyberbullying from images and videos, as visual forms of cyberbullying are perceived more harmful than the written ones by users [17]. This goal could be reached through the contribution of scholars from different fields, because of the technical (i.e., difficulty to create datasets containing this type of entries) and legal (i.e., privacy issues) issues raised by sharing multimedia contents. It is also necessary to understand which impact these detection systems could have on users' everyday life. Future works will be challenged to combine these technological systems with the implementation of psychosocial interventions. Therefore, a dialogue between social and computer sciences is essential to provide users with quick detection/prevention technology-based strategies and effective targeted interventions. This step will require the development of rigorous interdisciplinary research protocols, possibly based on Randomized Control Trials (RCT) studies. Finally, automatically detecting cyberbullying will help to avoid underestimating cyberbullying at school and to provide students, involved in any type of harassment, with timely interventions as traditional bullying dynamics and cyberbullying dynamics

are strongly interconnected [43]. For instance, app-based tools on this predictive modelling may represent a powerful resource for school institutions.

V.CONCLUSION

The present work has critically presented the strategies and the implications of ML for the automatic detection of cyberbullying. Ten literature reviews have been selected and discussed. Content-based features have been resulted the most used features in ML models predicting cyberbullying. Specifically, bag-of-words is the most common approach addressed to predict online bullying from users' online contents. As concerns alghoritms, Support Vector Machine (SVM), Naïve Bayes and Convolutional Neural Networks were shown to be the most performing algorithms, with SVM being the most efficient one. The practical implications of the use of these methods for cyberbullying prevention have been also analyzed. ML represents an innovative preventative strategy enabling fast and accurate detection of cyberbullying cases online. Thus, it may optimize and integrate psychoeducational programs for adolescents. However, there are limitations of this approach which rely on the lack of multimedia contents (e.g., images and videos) and of a shared definition of cyberbullying. Future research will be challenged to enhance the effectiveness of these algorithms, for instance, by training them to detect cyberbullying from several multimedia sources.

REFERENCES

1. D. Olweus, "Cyberbullying: An overrated phenomenon?" European Journal of Developmental Psychology 9, no. 5, 2012, pp. 520–538. https://doi.org/10.1080/17405629.2012.682358
2. A.C. Baldry, D.P. Farrington, A. Sorrentino, and C.Blaya, "Cyberbullying and cybervictimization" in International Perspectives on Cyberbullying. Cham: Palgrave Macmillan, 2018, pp. 3-23. https://doi.org/10.1007/978-3-319-73263-3_1
3. J.W. Patchin, and S. Hinduja, "Bullies move beyond the schoolyard: A preliminary look at cyberbullying." Youth Violence and Juvenile Justice 4 no. 2, 2006, pp. 148-169. https://doi.org/10.1177%2F1541204006286288
4. D. Olweus, and S.P. Limber, "Some problems with cyberbullying research." Current Opinion in Psychology 19, 2018, pp. 139–143. https://doi.org/10.1016/j.copsyc.2017.04.012
5. D. Olweus, "Bullying at school". In Aggressive behavior (pp. 97-130). Boston, MA: Springer, 1994, pp. 97-130.
6. F. Mishna, C. Cook, T. Gadalla, J. Daciuk, and S. Solomon, "Cyber bullying behaviors among middle and high school students." American Journal of Orthopsychiatry 80, no 3, 2010, pp. 362-374. https://doi.org/10.1111/j.1939-0025.2010.01040.x
7. D.L. Espelage, I.A. Rao, and R.G. Craven, "Theories of cyberbullying" in Principles of cyberbullying research, London: Routledge, 2012, pp. 77-95.
8. R. Agnew, and H.R. White, "An empirical test of general strain theory." Criminology 30 no 4, 1992, 475-500. https://doi.org/10.1111/j.1745-9125.1992.tb01113.x
9. L.E. Cohen, and M. Felson, "Social change and crime rate trends: A routine activity approach." American Sociological Review 44, no. 4, 1979, pp. 588-608. https://doi.org/10.2307/2094589
10. J. N. Navarro, and J.L. Jasinski, "Going cyber: Using routine activities theory to predict cyberbullying experiences." Sociological Spectrum 32,2012, pp. 81–94. https://doi.org/10.1080/02732173.2012.628560
11. U. Bronfenbrenner, "Toward an experimental ecology of human development." American Psychologist 32, 1977 pp. 513–531. https://doi.org/10.1037/0003-066X.32.7.513
12. M. Vyawahare, and M. Chatterjee, "Taxonomy of Cyberbullying Detection and Prediction Techniques in Online Social Networks". In Data Communication and Networks, Singapore: Springer, 2020, pp. 21-37. Springer, Singapore. https://doi.org/10.1007/978-981-15-0132-6_3
13. N. E. Willard, "Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress," Research Press, 2007.
14. P.K. Smith, C. del Barrio, and R.S. Tokunaga, "Definitions of Bullying and Cyberbullying: How Useful Are the Terms?". In Principles of cyberbullying research, 2012, London: Routledge, pp. 54-68.
15. D. Álvarez-García, A. Barreiro-Collazo, and J. C. Núñez, "Cyberaggression among adolescents: Prevalence and gender differences." Comunicar 25, 2017, pp. 89–97. https://doi.org/10.3916/C50-2017-08
16. P. Redmond, J. V. Lock, and V. Smart, "Developing a cyberbullying conceptual framework for educators." Technology in Society 60, 2019. https://doi.org/10.1016/j.techsoc.2019.101223
17. E. Menesini, A. Nocentini, and P. Calussi, "The measurement of cyberbullying: Dimensional structure and relative item severity and discrimination." Cyberpsychology, Behavior, and Social Networking 14, 2011, pp. 267–274. https://doi.org/10.1089/cyber.2010.0002
18. K. Hellfeldt, L. López-Romero, and H. Andershed, "Cyberbullying and psychological wellbeing in young adolescence: the potential protective mediation effects of social support from family, friends, and teachers." International Journal of Environmental Research and Public Health 17, no. 1, 2020, pp. 45. https://doi.org/10.3390/ijerph17010045
19. J. Wang, R. J. Iannotti, and J.W. Luk, "Peer victimization and academic adjustment among early adolescents: Moderation by gender and mediation by perceived classmate support." Journal of School Health 81, 2011, pp. 386–392. https://doi.org/10.1111/j.1746-1561.2011.00606.x
20. P.S. Gamito, D.G. Morais, J.G Oliveira, R. Brito, P.J. Rosa, and M.G. de Matos, "Frequency is not enough: Patterns of use associated with risk of Internet addiction in Portuguese adolescents." Computers in Human Behavior 58, 2016, pp. 471–478. https://doi.org/10.1016/j.chb.2016.01.013
21. S.M. Coyne, L.M. Padilla-Walker, H.G. Holmgren, E.J. Davis, K.M., Collier, M.K. Memmott-Elison, and A.J. Hawkins, "A meta-analysis of prosocial media on prosocial behavior, aggression, and empathic concern: A multidimensional approach." Developmental Psychology 54, 2018, pp. 331–347. https://doi.org/10.1037/dev0000412
22. C.P. Barlett, C.C. DeWitt, B. Maronna and, K. Johnson, "Social media use as a tool to facilitate or reduce cyberbullying perpetration: A review focusing on anonymous and nonanonymous social media platforms." Violence and Gender 5, 2018, pp. 147–152. https://doi.org/10.1089/vio.2017.0057
23. F.C. Chang, C.H., Chiu, N.F. Miao, P.H. Chen, C.M. Lee, T.F., Huang, and Y.C. Pan, "Online gaming and risks predict cyberbullying perpetration and victimization in adolescents." International Journal of Public Health 60, 2015, pp. 257–266. https://doi.org/10.1007/s00038-014-0643-x
24. Z. Hilvert-Bruce, and J.T. Neill, "I'm just trolling: The role of normative beliefs in aggressive behavior in online gaming." Computers in Human Behavior 102, 2020, pp. 303-311. https://doi.org/10.1016/j.chb.2019.09.003
25. Z. Ashktorab "The Continuum of Harm" taxonomy of cyberbullying mitigation and prevention. In Online Harassment, Springer: Cham, 2018, pp. 211-227. https://doi.org/10.1007/978-3-319-78583-7_9
26. A. L. Tarca, V.J. Carey, X.W. Chen, R. Romero, and S. Drăghici, "Machine learning and its applications to biology." PLoS Computational Biology 3, no. 6, 2007, e116.
27. https://doi.org/ 10.1371/journal.pcbi.0030116.
28. S. Navlakha, and Z. Bar-Joseph, "Algorithms in nature: the convergence of systems biology and computational thinking." Molecular Systems Biology 7, no. 1, 2011, pp. 546.
29. https://doi.org/ 10.1038/msb.2011.78
30. M.W. Libbrecht, and W. S. Noble, "Machine learning applications in genetics and genomics." Nature Reviews Genetics 16, no. 6, 2015, pp. 321-332. https://doi.org/10.1038/nrg3920
31. N. Navarin, and F. Costa, "An efficient graph kernel method for non-coding RNA functional prediction." Bioinformatics 33, no. 17, 2017, pp. 2642-2650. https://doi.org/ 10.1093/bioinformatics/btx295.
32. P. Sundsøy, J. Bjelland, A.M Iqbal, and Y.A. de Montjoye, "Big data-driven marketing: how machine learning outperforms marketers' gut-feeling," In 2014 International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, 2014, pp. 367-374. https://doi.org/10.1371/journal.pone.0073791

33. K. Siau, and Y. Yang, "Impact of artificial intelligence, robotics, and machine learning on sales and marketing," In 2017 Twelve Annual Midwest Association for Information Systems Conference (MWAIS), 2017, pp. 18-19.
34. G. Orru, W. Pettersson-Yeo, A.F. Marquand, G. Sartori, and A. Mechelli, "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review." Neuroscience Biobehavioral Reviews 36, no. 4, 2012, pp. 1140-1152. https://doi.org/ 10.1016/j.neubiorev.2012.01.004
35. M. Buscema, S.C. Ricerche, and E. Grossi, "The semantic connectivity map: An adapting self-organising knowledge discovery method in data bases." International Journal of Data Mining and Bioinformatics 2, no. 4, 2008, pp. 362-404. https://doi.org/10.1504/IJDMB.2008.022159.
36. L.G. Ahmad, A.T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, and A.R. Razavi, "Using three machine learning techniques for predicting breast cancer recurrence." Journal of Health and Medical Informatics 4, no. 124, 2013, pp. 3. https://doi.org/10.4172/2157-7420.1000124.
37. Z. Obermeyer, E.J., and Emanuel, "Predicting the future—big data, machine learning, and clinical medicine." The New England journal of medicine 375, no. 13, 2016, pp. 1216. https://doi.org/10.1056/NEJMp1606181
38. A. Borghesi, A. Bartolini, M. Lombardi, M. Milano, and L. Benini, "Anomaly detection using autoencoders in high performance computing systems."Proceedings of the AAAI Conference on Artificial Intelligence 33, 2019, pp. 9428-9433.
39. G. Orrù, M. Monaro, C. Conversano, A. Gemignani, and G. Sartori, "Machine learning in psychometrics and psychological research." Frontiers in Psychology 10, 2020, pp. 2970. https://doi.org/ 10.3389/fpsyg.2019.02970
40. A. Smola, and S. Vishwanathan, "Introduction To Machine Learning." Cambridge: Cambridge University Press, 2008.
41. C. Baldini, F. Ferro, N. Luciano, S. Bombardieri, and E. Grossi, "Artificial neural networks help to identify disease subsets and to predict lymphoma in primary Sjögren's syndrome." Clinical and Experimental Rheumatology 36, no. 112, 2018, pp. 137-144.
42. B. Kwolek, "Face detection using convolutional neural networks and Gabor filters." In 2005 International Conference on Artificial Neural Networks, Berlin: Springer, 2005, pp. 551-556.
43. https://doi.org/ 10.1007/11550822_8
44. G. Levi, and T. Hassner, "Age and gender classification using convolutional neural networks." In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 34-42.
45. K. N. Hampton, L.S. Goulet, L. Raine, and K. Purcell, "Social networking sites and our lives." Pew Internet and American Life Project, 2012.
46. J.W. Ybarra, and K.J. Mitchell, "Online aggressor/targets, aggressors, and targets: A comparison of associated youth characteristics." Journal of Child Psychology and Psychiatry 45, no. 7, 2004, pp. 1308-1316. https://doi.org/10.1111/j.1469-7610.2004.00328.x
47. R.P. Ang, "Cyberbullying: Its prevention and intervention strategies" in Child Safety, Welfare and Wellbeing. New Delhi: Springer, 2016, pp. 25-38.
48. Q. Li, "Cyberbullying in high schools: A study of students' behaviors and beliefs about this new phenomenon." Journal of Aggression, Maltreatment Trauma 19, no. 4, 2010, pp. 372-392. https://doi.org/10.1080/10926771003788979
49. H. Dani, J. Li, and H. Liu, "Sentiment informed cyberbullying detection in social media." In 2017 Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Cham: Springer, 2017, pp. 52-67. https://doi.org/ 10.1007/978-3. 2017
50. Q. Huang, V.K. Singh, and P.K. Atrey, "Cyber bullying detection using social and textual analysis." In 2014 Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, 2014, pp. 3-6. https://doi.org/10.1145/2661126.2661133
51. M. Ptaszynski, F. Masui, Y. Nakajima, Y. Kimura, R. Rzepka, and K. Araki, "Detecting cyberbullying with morphosemantic patterns." In 2016 Proceedings of the Joint 8th International Conference on Soft Computing and Intelligent Systems and 17th International Symposium on Advanced Intelligent Systems (SCIS-ISIS), 2016, pp. 248-255.
52. M. Dadvar, F.D. Jong, R Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information." In Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR), 2012. University of Ghent.
53. S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey." IEEE Transactions on Affective Computing 11, no. 1, 2017, pp 3-24. https://doi.org/10.1109/TAFFC.2017.2761757
54. A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network." In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015, pp. 280-285. https://doi.org/10.1145/2808797.2809398
55. N. Potha, and M. Maragoudakis, "Cyberbullying detection using time series modeling." In 2014 IEEE International Conference on Data Mining Workshop, 2014, pp. 373-382. https://doi.org/10.1109/ICDMW.2014.170
56. M. Petticrew, and H. Roberts, "Systematic reviews in the social sciences: A practical guide." Hoboken: John Wiley Sons, 2008.
57. C.M. Bishop, "Pattern recognition and machine learning". Springer: Singapore, 2008
58. M.A. Al-Garadi, M.R. Hussain, N. Khan, G. Murtaza, H.F. Nweke, G. Ali Mujtaba, H. Chiroma, H. Ali Khattaki, and A. Gani, "Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges." IEEE Access 7, 2019, pp 70701–70718.
59. https://doi.org/ 10.1109/ACCESS.2019.2918354
60. S. Nadali, M. Azrifah, A. Murad, N. Bt, M. Sharef, A. Mustapha, and H. Onn, "A review of cyberbullying detection: An Overview." In 2013 13th International Conference on Intelligent Systems Design and Applications, 2013, pp. 325-330. https://doi.org/10.1109/ISDA.2013.6920758
61. A. Singh, and M. Kaur, "Content-based cybercrime detection: A concise review." International Journal of Innovative Technology and Exploring Engineering 8, no. 8, 2019, pp. 1193–1207. https://doi.org/ 10.1007/s11227-019-03113-z
62. U. Can, and B. Alatas, "A new direction in social network analysis: Online social network analysis problems and applications." Physica A: Statistical Mechanics and Its Applications 535, 2019, 1-38. https://doi.org/10.1016/j.physa.2019.122372
63. R.M. Kowalski, and S.P. Limber, "Electronic bullying among middle school students." Journal of Adolescent Health, 41, 2007, pp. 22–S30. https://doi.org/10.1016/j.jadohealth.2007.08.017
64. T. Mahlangu, C. Tu, P. Owolawi, and A.T. Definition, "A Review of Automated Detection Methods for Cyberbullying." 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC), 2018, 1–5.
65. J.F. Chisholm, "Cyberspace violence against girls and adolescent females." Annals of the New York Academy of Sciences 1087, no. 1, 2006, pp. 74-89. https://doi.org/10.1196/annals.1385.022.
66. G. Perasso, and L. Barone, "Cyberbullismo, cyber-vittimizzazione e differenze di genere in adolescenza." Psicologia Clinica Dello Sviluppo 22, no. 2, 2018, pp. 241–268. https://doi.org/10.1449/90830
67. I. Ajzen "The theory of planned behavior." Organizational Behavior and Human Decision Processes 50, no. 2, 1991, pp. 179-211.
68. H. Rosa, N. Pereira, R. Ribeiro, P.C. Ferreira, J.P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A.M. Veiga Simao, and I. Trancoso, "Automatic cyberbullying detection: A systematic review." Computers in Human Behavior 93, 2019, pp. 333-345. https://doi.org/10.1016/j.chb.2018.12.021
69. T. Evgeniou, and M. Pontil, "Support vector machines: Theory and applications." In Advanced Course on Artificial Intelligence. Berlin: Springer, 1999, pp. 249-257. https://doi.org/ 10.1007/3-540-44673-7_12
70. S.F. Chan, A.M. La Greca, and J.L. Peugh, "Cyber victimization, cyber aggression, and adolescent alcohol use: Short-term prospective and reciprocal associations." Journal of adolescence 74, 2019, pp. 13-23. https://doi.org/10.1016/j.adolescence.2019.05.003.
71. R. Del Rey, R. Ortega-Ruiz, and J.A. Casas, "Asegúrate: An intervention program against cyberbullying based on teachers' commitment and on design of its instructional materials." International journal of environmental research and public health 16, no. 3, 2019, pp. 434. https://doi.org/10.3390/ijerph16030434.
72. E. Menesini, A. Nocentini, B.E. Palladino, A. Frisén, S. Berne, R. Ortega-Ruiz, J. Calmaestra, H. Scheithauer, A. Schultze-Krumbholz, P. Luik, K. Naruskov, C. Blaya, J. Berthaud, and P.K. Smith. "Cyberbullying definition among adolescents: A comparison across six European countries." Cyberpsychology, Behavior, and Social Networking 15, no. 9, 2012, pp. 455-463. https://doi.org/10.1089/cyber.2012.0040.
73. D.H. Wolpert, and W.G. Macready, "No free lunch theorems for optimization." IEEE Transactions on Evolutionary Computation 1, no. 1, 1997, pp. 67-82 https://doi.org/10.1109/4235.585893

74. A. Ashikin, and S. Norhalina, "A Review on Violence Video Classification Using Convolutional Neural Networks." International Conference on Soft Computing and Data Mining, Cham: Springer, 2016. https://doi.org/10.1007/978-3-319-51281-5_14.
75. R. Trevethan, "Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice", Frontiers in Public Health 5, 2017. Available: 10.3389/fpubh.2017.00307. https://doi.org/10.3389/fpubh.2017.00307.
76. M. Kuhn, and K. Johnson, "Applied predictive modeling." 26, New York: Springer, 2013.
77. C. Blaya, "Aggression and Violent Behavior Cyberhate: A review and content analysis of intervention strategies." Aggression and Violent Behavior 45, 2019, pp. 163–172.
78. https://doi.org/ 10.1016/j.avb.2018.05.006
79. M. Fire, R. Goldschmidt, and Y. Elovici, "Online Social Networks: Threats and Solutions." IEEE Communications Surveys & Tutorials 16, no. 4, 2019, pp. 2019–2036.
80. https: //doi.org/10.1109/COMST.2014.2321628
81. A. Nocentini, V. Zambuto, and E. Menesini, "Aggression and Violent Behavior Anti-bullying programs and Information and Communication Technologies (ICTs): A systematic review." Aggression and Violent Behavior 23, 2015, pp. 52–60. https://doi.org/10.1016/j.avb.2015.05.012
82. P.J.C. Pérez, C.J.L.Valdez, M.D.G.C. Ortiz, J.P.S. Barrera, and P.F. Pérez, "Misaac: Instant messaging tool for ciberbullying detection." In Proceedings on the International Conference on Artificial Intelligence (ICAI), 2012, pp. 1-4.
83. M.L. Mitchell, and J.M. Jolley, "Research design explained." Cengage Learning, 2012.
84. J.L. Castroman, B. Moulahi, J. Azé, S. Bringay, J. Deninotti, S. Guillaume, and E. Baca-Garcia, "Mining social networks to improve suicide prevention : A scoping review", Journal of Neuroscience Research 98, no.4, 2020, pp. 616-625. https://doi.org/10.1002/jnr.24404
85. N. Chetty, and S. Alathur, "Aggression and violent behavior hate speech review in the context of online social networks." Aggression and Violent Behavior 40, 2018, pp. 108–118. https://doi.org/10.1016/j.avb.2018.05.003
86. M Hudan, N. Sunan, K. Yogyakarta, U.P. Bandung, U. Sultan, Z. Abidin, and M. Lampung, "From live interaction to virtual interaction: addressing moral engagement in the digital era" Journal of Theoretical and Applied Information Technology 95, no. 19, 2017, pp. 4964-4972.
87. M. Lancaster, "A systematic research synthesis on cyberbullying interventions in the United States. "Cyberpsychology, Behavior, and Social Networking 21, no. 10, 2018, pp. 593-602. https://doi.org/10.1089/cyber.2018.0307
88. N. R. Nugent, S. R. Pendse, H.T. Schatten, and M. F. Armey, "Innovations in Technology and Mechanisms of Change in Behavioral Interventions." Behavior Modification, 2019 pp. 1-29. https://doi.org/10.1177%2F0145445519845603
89. A. R. Pisani, P.A. Wyman, D. C. Mohr, T. Perrino, C. Gallo, J. Villamar, K. Kendziora, G.W. Howe, Z. Sloboda, and C. H. Brown, "Human Subjects Protection and Technology in Prevention Science: Selected Opportunities and Challenges." Prevention Science, 17, no. 6, 2016, pp. 765–778. https://doi.org/10.1007/s11121-016-0664-1
90. R. J. Watson, and J. L. Christensen, "ScienceDirect Big data and student engagement among vulnerable youth: A review." Current opinion in behavioral sciences 18, 2017, pp. 23-27. https://doi.org/10.1016/j.cobeha.2017.07.004

AUTHORS PROFILE

Jacopo De Angelis, is PhD Candidate at University of Milano-Bicocca (Italy). After obtaining a MSc in experimental psychology and cognitive neuroscience at University of Pavia (Italy), he spent six months in the UK working as a research intern at Goldsmiths University of London, and six months working at the Italian Institute of Technology (Genoa, Italy) as a junior research fellow. His main research interest concerns the study of cognitive and emotional functions of the brain such as attention and emotion, as well as their applications in clinical sciences, business, and technology. His approach combines traditional statistical methods with more advanced computational and data science approaches (i.e., machine learning, data mining, and deep learning). Accordingly, he is also interested in studying how advanced predictive modelling can contribute to i) developing models of cognitive processes, with a particular focus on attentional, decision making and sensory processes; ii) building predictive aetiologic, diagnostic and intervention models of different neuropsychiatric pathologies (e.g., ASD); iii) predicting behaviors and attitudes in several contexts (e.g., consumer behavior and choice, cyberbullying, marketing etc.). He keeps upgrading his knowledge by attending (both as presenter and attendee) several specialization courses, conferences, workshops and summer schools, especially in the field of cognitive neurosciences, data science, statistics and machine learning.

Giulia Perasso, is a PhD Candidate in Psychology, Neuroscience and Data Science at University of Pavia (Italy). After obtaining a MSc in Psychology (2015), she researched as visiting intern at Middlesex University (London, UK), approaching the theme of juvenile cybercrime (2016). In 2017, she obtained another MSc degree in Criminology at Sapienza University (Rome, Italy). Nowadays, she is researching about cyberbullying victimization during adolescence, analyzing the epidemiological dataset of the Health Beahviour in School-Aged Children protocol (data collection from Lombardy, Italy), promoted by World Health Organization. Cyberbullying predictors, gender differences in cyberbullying, and cyberbullying prevention constitute the focus of her research and publications.

Appendix A: Table displaying data-extraction tool and inter-rater evaluation.

Studies	RATERS	Theoretical Background (total=9)			Research questions and aims (total=6)		Search strategy and Method (total=12)				Results (total=9)			TOTAL		Inclusion
		appropriateness of citations 1-3	accuracy: core topic 1-3	scientific debate (innovative) 1-3	A priori =1 or explorative =0	Clarity 1-5	Systematic review (Y=2, N=0)	Synonyms (2= good, 1= reported filter not specific, 0= no)	Number of REPORTED datasets 0-5	Inclusion/exclusion criteria (0=no details, 3=detailed)	Impact (1-5)	Details (0=no, 1=yes, 2=good)	Limits discussed (0=no, 1=yes, 2=good)	TOT	RATERS TOT	
Blaya [73]	R1	2	2	2	1	5	2	2	5	3	3	2	0	54	29	Included
	R2	1	1	1	1	5	2	2	5	3	2	2	0		25	
Ashikin & Norhalina [70]	R1	3	3	3	0	5	0	0	0	0	4	1	0	36	19	included
	R2	3	3	3	0	3	0	0	0	0	5	0	0		17	

Al-Garadi et al. [55]	R1	2	3	3	0	5	2	2	5	1	5	2	0	61	30	Included
	R2	2	3	3	0	5	2	2	5	1	5	2	1		31	
Can & Alatas [58]	R1	2	3	3	0	4	0	0	0	0	5	2	0	40	19	included
	R2	3	3	3	0	5	0	0	0	0	5	2	0		21	
Castroman et al. [78]	R1	2	2	1	0	1	0	2	4	0	4	2	0	31	18	
	R2	2	1	1	0	1	0	2	4	0	1	1	0		13	
Chetty & Alathur [79]	R1	2	2	2	0	2	0	0	0	0	3	1	0	20	12	
	R2	2	1	2	0	1	0	0	0	0	2	0	0		8	
Fire et al. [74]	R1	2	2	2	0	4	0	0	0	0	5	2	0	39	17	included
	R2	3	3	3	0	5	0	0	0	0	5	2	1		22	
Huda et al. [80]	R1	2	1	1	0	3	0	1	0	0	3	1	0	23	12	
	R2	2	2	1	0	3	0	1	0	0	2	0	0		11	
Lancaster [81]	R1	0	0	1	0	2	2	2	3	2	4	1	0	30	17	
	R2	0	0	0	0	2	2	2	3	1	1	2	0		13	
Mahlangu et al. [60]	R1	2	3	2	0	5	1	0	5	0	4	2	1	52	25	included
	R2	2	3	3	0	5	1	0	5	0	5	2	1		27	
Nadali et al. [56]	R1	3	3	3	0	2	0	0	0	0	3	1	0	36	15	included
	R2	3	3	3	0	5	0	0	0	0	5	1	1		21	
Nocentini, et al. [75]	R1	2	2	2	0	3	2	2	3	3	3	1	2	51	25	included
	R2	2	2	3	0	4	2	2	3	3	3	1	1		26	
Nugent et al. [82]	R1	1	1	2	0	1	0	0	0	0	1	1	0	14	7	
	R2	1	1	1	0	2	0	0	0	0	2	0	0		7	
Pisani et al. [83]	R1	1	2	1	0	1	0	0	0	0	2	1	0	14	8	
	R2	1	1	1	0	1	0	0	0	0	1	1	0		6	
Rosa et al. [64]	R1	2	3	3	1	5	2	0	5	0	4	2	2	59	29	included
	R2	3	3	3	1	5	2	0	5	0	5	1	2		30	

Singh & Kaur [57]	R1	2	3	3	0	4	0	1	0	0	4	2	1	43	20	Included
	R2	3	3	3	0	5	0	1	0	0	5	2	1		23	
Watson & Christensen [84]	R1	1	2	2	0	2	0	0	0	0	5	1	1	31	14	
	R2	2	2	2	0	3	0	0	0	0	4	2	2		17	

Note: The quality assessment included an evaluation from 1 to 9 for the theoretical background (1 to 3 points for citations' appropriateness; 1 to 3 points for core-topic accuracy; 1 to 3 points for innovation), an evaluation of 1 to 6 for the research questions and aims (1 point for a priori hypothesis, 0 for explorative; 1 to 5 points for clarity of the aims), an evaluation from 1 to 12 for the search strategy and method (0 points for non-systematic reviews, 1 point for semi-systematic reviews, 2 points for systematic reviews; 0 to 5 points as the equivalent of scientific databanks consulted; 0 to 3 points for the accuracy of inclusion and exclusion criteria), an evaluation from 1 to 9 of the results part (1 to 5 points for the impact of the reviews; 0 to 2 points for detailedness; 0 to 2 points for review limits discussion). By summing the two raters' evaluations, the maximum achievable score was 72. The inclusion cut-off was set at 36 points.