Cyberbullying detection - A Systematic Literature Review of AI-Based Solutions

Raquel Moreira Machado Fernandes, Programa de Pós-Graduação em Informática - Universidade Federal do Rio de Janeiro, raquel.fernandes@ufrj.br, https://orcid.org/0000-0001-7643-6925
Luiz Fernando Rust da Costa Carmo, Programa de Pós-Graduação em Informática - Universidade Federal do Rio de Janeiro, rust@nce.ufrj.br, https://orcid.org/0000-0001-6131-7771
Claudia Lage Rebello da Motta, Programa de Pós-Graduação em Informática - Universidade Federal do Rio de Janeiro, claudiam@nce.ufrj.br, https://orcid.org/0000-0002-4069-1462

ABSTRACT
This exploratory-descriptive research used the systematic literature review method to examine the current state of Cyberbullying detection using Artificial Intelligence. We analyzed 34 papers from 6 databases, highlighting techniques, algorithms and datasets that have been used, along with practical applications developed in the last five years. Additionally, we identified research gaps and opportunities to propose mitigating measures against cyberbullying in Brazilian primary and secondary school environments, as well as the development of a browser extension to detect and notify the writing of potentially offensive messages.

Keywords: *Cyberbullying, Detection, Artificial intelligence.*

Detecção de cyberbullying - Uma revisão sistemática da literatura de soluções baseadas em IA

RESUMO
Esta pesquisa exploratório-descritiva utilizou o método de revisão sistemática da literatura para examinar o estado atual da detecção de Cyberbullying com a utilização de Inteligência Artificial. Analisamos 34 artigos provenientes de 6 bases de dados, destacando técnicas, algoritmos e conjuntos de dados que vêm sendo utilizados, juntamente com aplicações práticas desenvolvidas nos últimos cinco anos. Além disso, identificamos lacunas e oportunidades de pesquisa para propor medidas mitigadoras contra o cyberbullying em ambientes escolares de educação básica brasileiros, bem como o desenvolvimento de uma extensão de navegador para detectar e notificar a escrita de potenciais mensagens ofensivas.

Palavras-chave: *Cyberbullying, Detecção, Inteligência Artificial.*

1 Introduction

Cyberbullying has been continually listed as a major concern regarding children's use of the Internet (United Nations, 2022; UNESCO, 2023; Security.Org, 2024). This practice has been reported as a growing threat to global public health. There are many types of cyberbullying (Bauman, 2015) and several studies already highlighted the relationship between cyberbullying and damage to the mental health of children and adolescents. In addition to problems such as

depression, anxiety and conduct disorder as highlighted by Evans-Lacko et al. (2017), cyberbullying can increase the risk of suicide or suicide-related behaviors, such as suicidal ideation, suicide attempt, and self-harm (Price & Dalgleish, 2010).

In this scenario, the search for solutions to identify and prevent cyberbullying has been a priority for researchers, educators and professionals from different areas of knowledge. In this context, artificial intelligence (AI) has emerged as a promising tool to assist in detecting and mitigating cyberbullying. AI's ability to analyze large volumes of data in real time, identify patterns and predict behavior offers opportunities to tackle this problem.

The present study aimed to investigate the application of AI techniques in detecting cyberbullying, in order to understand the current panorama of research and identify opportunities and possible gaps in this area. The study is structured into 4 sections, the first of which is the presentation of the topic. In section 2 we present the methodological approach of the research. In section 3 we present results and discussion, and in section 4 we present final considerations, gaps, limitations and future perspectives.

2 Method

This exploratory-descriptive research utilized the Systematic Literature Review (SLR) method, following the guidelines outlined by Dermeval, Coelho and Bittencourt (2020). For research protocol development, we followed Kitchenham and Charters (2007). This methodology was selected due to its ability to provide a comprehensive and structured view of the current state of knowledge in a given area, in accordance with the objective of this research.

This SLR seeks to answer three main research questions (RQ) related to the use of AI to detect cyberbullying, specifically: RQ1. What AI techniques have been used to detect cyberbullying? RQ2. What has been used to construct the training datasets? RQ3. What applications have been developed in the last 5 years using AI with the aim of detecting cyberbullying?

The research was conducted through the analysis of studies from 6 databases, namely: i) Scopus; ii) IEEE; 3) CAPES periodicals; 4) Google Scholar; 5) Scielo and 6) SBC OpenLib. As search strings, we used: "cyberbullying", "artificial intelligence" and "detection", with the aim of expanding the scope of the research, allowing us to capture a variety of studies, techniques and approaches related to these topics.

Papers were retrieved in November/December 2023. In the first search, 206 candidate papers were found throughout all databases. As inclusion criteria we selected: IC1. Publications that report an experience to detect cyberbullying using artificial intelligence techniques; IC2. Publications in the last 5 years, considering the period from 2018 to 2023.

As exclusion criteria we selected: EC1. Publications where the full text was not available; EC2. Publications not found in journals or scientific conferences; EC3. Publications that address the topic of cyberbullying, but do not present an artificial intelligence strategy of detection; EC4. Publications that do not present an algorithm test, or a detection proposal, such as literature review.

3 Results and Discussion

Following manual screening of 206 candidate papers against inclusion and exclusion criteria, the authors of this review included 34 studies for analysis, with the following approximate

distribution: Periodicals CAPES (47.06%), Scopus (35.29%), IEEE (14.71%) and Google Scholar (2.94%). Several significant research points were identified, including:

3.1 Detection techniques with AI-Based Solutions

This section responds to RQ1 and presents artificial intelligence approaches that have been used to detect Cyberbullying. In most of the studies, machine learning techniques were employed for cyberbullying detection, being presented as the most satisfactory approach. According to Alduailaj and Belghith (2023), machine learning (ML) is a subfield of AI that offers systems with the capability of learning and improving with the involvement of automation processes from previous experience and without having to be specifically programmed were a system learn for itself to make decisions instantly by applying a training dataset.

Some studies like that of Paruchuri and Rajesh (2022) used Deep learning (DL) which is a subset of Machine Learning (ML). This technique is based on artificial neural networks composed of multiple layers of interconnected neurons, and unlike traditional machine learning models, DL allows the model to learn automatically from training examples.

Neural networks are computational models inspired by the functioning of the human brain, composed of layers of interconnected neurons that process information and learn from examples. These models were also used. For example, Murshed et al. (2022) used Elman type Recurrent Neural Networks (RNN), a recurrent neural network technique capable of learning and processing sequences of data while maintaining internal memory of recent information. Another variation is called convolutional neural networks (CNN), experimented by Roy and Mali (2022).

Natural Language Processing (NLP) is another technique identified in some studies like Dewani et al. (2023), who used NLP to treat Roman Urdu micro text. NLP involves the development and application of algorithms, models and methods to process and understand human language in an automated way. In this way, it is possible to extract meaning and useful information from texts. Notice that ML provides the tools and techniques needed to perform NLP tasks.

We also find, as in Teng and Varathan (2023) and Roy and Mali (2022), the use of Transfer Learning. This technique involves reusing knowledge acquired by a model trained for a specific task, such as text classification, to improve performance on a related task, such as cyberbullying detection.

Still in the field of machine learning, in Daniel et al. (2023) and Azeez et al. (2021) we find the use of ensemble learning process, a technique that combining multiple machine learning classifiers and models, attempting to produce better results than constituent models. Another technique is sentiment analysis, as used by Al-Hashedi et al. (2023). This technique understanding information about emotions, opinions, behaviors and attitudes through texts. Al-Hashedi et al. (2023) proposes cyberbullying detection models that are trained based on contextual, emotions and sentiment features. For them, anger, fear and guilt were the major emotions associated with cyberbullying.

Balakrishnan, Khan, and Arabnia (2020) uses psychological features, including personality, sentiment, and emotion, and show that cyberbullying detection improved when personalities and sentiments were used. However, the authors did not observe a similar effect for emotion.

Among the 34 papers examined, only one paper focused on addressing image-based cyberbullying incidents. Roy & Mali (2022) presents that cyberbullying with the image received comparatively less attention, and this represents a research gap. The authors proposed a model using Deep learning based Convolutional Neural Network (CNN) and transfer learning to predict image-based cyberbullying posts.

3.1.1 Algorithms and methods

The analyzed studies proposed algorithms, as Wu et al. (2023), or presented comparisons between different algorithms and existing approaches as Barthi et al. (2022). We found that the five most used algorithms were Support Vector Machines (SVM), followed by Random Forest (RF), Naïve Bayes, Linear Regression (LR) and Logistic Regression.

According to Chia et al. (2021), SVM is a supervised learning algorithm for classification or regression that uses the kernel trick to find an optimal boundary between the possible output.

Saravanaraj et al. (2016) apud Balakrishnan, Khan and Arabnia(2020) defines Random Forest as an ensemble learning algorithm that generates multiple small decision trees (or forests) from random subsets of features, each capturing different trends in the data. This algorithm performed better in many works such as Thun, Teh and Cheng (2022) and Azeez et al. (2021)

According Eronen et al. (2021), NaïveBayes (NB) classifier is a supervised learning algorithm applying Bayes' theorem that has a strong (naïve) assumption of independence between pairs of features. We also find variations of Naïve Bayes, such as Gaussian Naïve Bayes, Complement Naïve Bayes and Naïve Bayes (Bernouli) in Thun, Teh and Cheng (2022).

The linear regression (LR) algorithm is a supervised learning technique that models the relationship between a continuous output variable and one or more input variables, assuming a linear relationship among them.

According to Azeez et al. (2021), Logistic Regression is a widely used machine learning classifier that measures the relationship between independent variables (features) and a dependent variable (labels) by calculating probabilities using the sigmoid function.

Some other algorithms used were: ANN (Artificial Neural Networks), AdaBoost (AB), Extra Tree (ET), Linear Support Vector Classification (SVC), K-Nearest Neighbors (kNN), JRIP and J48.

Zambrano et al. (2023) used Latent Dirichlet Allocation (LDA) for topic modeling. Topic modeling is an unsupervised machine learning technique, which is part of the concept of AI and specifically NLP. This type of algorithm allows discovering topics (i.e. a set of terms) in a collection of documents.

Word embedding is a natural language processing methodology for mapping words or phrases from a lexicon to a corresponding vector of real numbers, which is then used to find word predictions and semantics (Raj et al., 2022). This technique and variations such as GloVe6B, GloVe42B, GloVe840B) were used e.g by Barthi et al.(2022), Rasool et al. (2023) and Raj et al. (2022), that used it to solve various issues that the simple one-hot vector encodings have.

Song et al. (2020) used text mining, a method for extracting information and patterns from large volumes of text. The authors employed methods to compute Term Frequency (TF), Document Frequency (DF), Degree of Visibility (DoV), and Degree of Diffusion (DoD), as well as to visualize

results through Keyword Emergence Map (KEM) and Keyword Issue Map (KIM) in a study in South Korea.

Daniel et al. (2023) used ensemble Deep Learning with Tournament-Selected Glowworm Swarm Optimization (EDL-TSGSO) algorithm and Ensemble Long Short-Term Memory with Adaboost (ELSTM-AB) model. In their proposal, the ensemble ELSTM-AB classifier integrates the prediction of LSTM and Adaboost models to enhance the overall classification performance.

According to Swetha and Ravikumar (2017), fuzzy logic is based on "degrees of truth" rather than the binary Boolean logic. Obaid, Guirguis, and Elkaffas (2023) used fuzzy logic to classify comments by severity, assigning them as Neither, High Bullying or Low Bullying.

Al-Marghilani (2022) proposed a detection model that primarily encompasses pre-processing and feature extraction processes. The technique involves chaotic salp swarm optimization (CSSO)-based feature selection technique to derive a useful set of features from the data. The author made an application of CSSO algorithm to choose optimum features and mayfly optimization (MFO) algorithm to tune the parameters.

Wu et al. (2023) works from the perspective of text classification and used capsule networks, that is, a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity.

Ptaszynski et al. (2019) demonstrates the application of exhaustive search algorithms in their research. The authors explain that such algorithms initially generate a massive number of combinations or potential answers to a given problem, which is why they are sometimes referred to as brute-force search algorithms.

Sangwan and Bathia (2020) investigated the correlation between denigration bullying and rumors and related the use of The Wolf Search Algorithm (WSA) which is formulated by simulation of the preying behavior of wolves. According to the authors, WSA can be visualized as multiple individual wolves gathering from various directions towards the optimal solution, instead of a single herd searching for the best solution in one direction at a time.

We observed distinct approaches among supervised, semi-supervised, and unsupervised techniques. Most commonly used was supervised learning, which involves training a model using a set of labeled data. Unsupervised learning involves identifying patterns or structures in data without the need for prior labels, while semi-supervised learning combines elements of both, using both labeled and unlabeled data to train a model.

3.2 Datasets

With respect to the datasets used in cyberbullying detection research, most of the datasets were constructed from Twitter, currently X[1]. One possible explanation for the wide use of Twitter is the fact that it has an API, as reported by Sainju et al. (2021) and Al-Hashedi et al. (2023) among others, which makes it easier for researchers and developers to access data. However, the use of Twitter data may not be representative in some cases. In Brazil, for example, in 2023, other social media platforms were among the most popular, such as Instagram[2] (89.8%), Facebook[3] (86.8%) and

[1] https://x.com/
[2] https://www.instagram.com/
[3] https://www.facebook.com/

TikTok[4] (65.9%) according to data from the Survey Digital (Data Reportal, 2023). Furthermore, we emphasize that solutions for cyberbullying must consider the target audience. Adults can also be victims of cyberbullying. However, this problem is more common in children and adolescents. A recent study by the World Health Organization (WHO, 2024) revealed that one in six children aged 11 to 15 has suffered cyberbullying. Still exemplifying the case of Brazil and according to the TIC Kids Online Brazil 2023 Survey (CETIC, n.d.), the most used networks are Instagram (preferred by teenagers), YouTube[5] (most accessed by younger people) and TikTok.

Of all the datasets analyzed, only that of Song et al. (2020) explicitly mentioned the use of terms related to children and adolescents. In this way, we emphasize the need for analyzes based on other datasets, which can provide more representativeness through the adequate capture of diversity from a faithful sample of the population and the phenomenon to be analyzed, so that the analyses can be generalized with greater confidence to a broader context.

We also consider that the datasets used may be a research gap, as we did not identify a standard or method of collection, cleaning and annotation, and these processes were carried out in different ways. We also emphasize that apart from Ptaszynski et al. (2019), who utilized a dataset compiled from data provided by official sources, we did not encounter similar initiatives. Ptaszynski et al. (2019) used a dataset created with data provided by the Human Rights Research Institute Against All Forms for Discrimination and Racism in Mie Prefecture, in Japan. This study mentioned the use of official instructions included in the MEXT manual for dealing with cyberbullying. This gap concerns the way in which the problem of cyberbullying has been treated and observed by governments, which demands the existence of public policies and official mapping of cyberbullying records, which could assist detection research.

Another issue identified regarding the datasets was that some annotations were carried out by untrained laypeople, as pointed out by Eronen et al. (2021) in relation to the Kaggle Formspring Dataset for Cyberbullying Detection (Reynolds et al., 2011) used in one of their four experiments.

The majority of the analyzed studies used a dataset consisting of 80% for training and 20% for testing, although some have used a different proportion. Some authors also used a portion for validation, such as Wu et al. (2023), which randomly divides the dataset into training, validation, and test sets using a ratio of 70:10:20.

To deal with unbalanced amounts of cyberbullying and non-cyberbullying data, Rasool et al. (2023) and Murshed et al. (2022), among others, mentioned the use of Synthetic Minority Oversampling Technique (SMOTE).

Note that not all papers mention the country where the research was carried out or the language used to construct the datasets. Furthermore, many papers do not describe the links or profiles from which the data was collected, nor the extraction criteria.

To examine RQ2, we present data on the datasets that have been used in the Table 1.

Table 1: Composition of Training Datasets Utilized in Cyberbullying Detection with AI

[4] https://www.tiktok.com/
[5] https://www.youtube.com/

AUTHORS	SOCIAL MEDIA PLATFORM	LANGUAGE	CLASSIFICATION	DATASET SIZE	DATA ANNOTATION
YUVARAJ ET AL. (2021)	Twitter	Not specified	CB and non-CB	30,384 tweets containing both CB and non-CB tweets	Unsupervised learning without manual labeling or tagging
LÓPEZ-VIZCAÍNO ET AL. (2021)	Vine social network	Not specified	Cyberbullying or normal	77,461 comments (16,332 cyberbullying / 61,129 normal) and others items such as media sessions, likes, among others	It was labeled as cyberbullying or normal using crowdsourcing
CHIA ET AL. (2021)	(i)A dataset provided by Semantic Evaluation 2018 Task 3: Irony Detection in English Tweets (Van Hee et al., 2018) (ii) A dataset collected by Ghosh and Veale (2017) (iii) A dataset from the Formspring data which consists of 12,728 data samples.	English	Ironic and non-ironic	i)4618 tweets (2222 ironic and 2396 non-ironic) (ii)51,189 tweets (24,453 sarcastic tweets and 26,736 non-sarcastic tweets) (iii) 12,728 data samples	(i)Manually labeled by three students using the Brat Rapid annotation tool2 with an inter-annotator agreement study (ii) Automatically collected from Twitter using user's self-declaration of sarcasm/irony with sarcastic and ironic hashtags (e.g. #irony, #sarcasm) and annotated for confirmation. (iii) Manual annotation with highly trained data annotators
BALAKRISHNAN, KHAN AND ARABNIA (2020)	Twitter	English	Bullies, spammers, aggressors, normal	5453 tweets (i.e., normal = 3510, 64.3%; spammer = 1489, 27.3%; aggressor = 173, 3.1%; bully = 281, 5.1%)	Annotation via crowdsourcing in which 834 contributors were recruited from various countries including the United States, Venezuela and Russia
THUN, TEH AND CHENG (2022)	Twitter	English	1 - Cyberbullying 0 - Normal tweet	5000 tweets (After eliminating all unwanted tweets, a total of 1200 tweets were acquired whereby 1000 were related to non-cyberbullying and 200 were related to cyberbullying)	Manually reviewed by the developer.
BALAKRISHNAN ET AL. (2019)	Twitter	Not specified	Bullies, spammers, aggressors, normal	9484 tweets, labeled as spammers (N = 3208; % = 33.8), aggressors (N = 336; % = 3.5), bullies (N = 528; % = 5.6) and normal (N = 5608; % = 59%).	Manually annotated data.
ERONEN ET AL. (2021)	(i) Kaggle Formspring dataset for cyberbullying detection as english cyberbullying (ii) A dataset for Japanese cyberbullying (iii)A Polish dataset (iv) A verification dataset, a subset of Yelp's user reviews data.	English, japanese, Polish	Harmful (CB) and non-harmful (non-CB)	(i) 300 thousand of tokens (ii) 1,490 harmful and 1,508 non-harmful entries (iii) 11,041 entries in total, with 10,041 included in the training set and 1000 in the test set. (iv) 250,000 positive and 250,000 negative reviews assigning one star reviews to negative class and five star reviews to positive class.	(i) Experts with psychological backgrounds re-annotated the dataset. (ii) Internet Patrol members labeled entries with the help of the Ministry of education, culture, sports, science and technology (MEXT) supplied manual (MEXT, 2008) (iii) Laypeople initially annotated, then corrected by an expert if needed. (iv) Not specified.
SONG AND SONG (2021)	Twitter	Korean	Perpetrator, victim, bystander	435,563 cases of cyberbullying-related topics	Using an opinion-mining method, the types of cyberbullying were sorted using SPSS 25.0.
SAINJU ET AL. (2021)	Twitter	English	Bullying trace and non-bullying trace	7868 tweets within the 7868 labeled tweets, 3096 or 39.35% were labeled as bullying traces.	Two of the study authors' labeled 7868 randomly selected tweets from the data collection period.
SANGWAN AND BATHIA (2020)	Comments posted on Instagram photos and tweets of global celebrities and world leaders/politicians considered rumourous	Not specified	600 'reputation rumours' and 600 'non-reputation rumours'	1200 comments	Comments are reviewed and identified manually.
ZAMBRANO ET AL. (2023)	Chats (short text data) or experiences of victims who have suffered some type of cyberbullying from Facebook	Not specified	Not specified	250,000 attacker related tweets / 3035 victim experiences	The Bullying data was analyzed by the WEB EMPATH tool, which automatically classified the words into lexical categories.
SONG ET AL. (2020)	Web documents from 279 subject channels	Korean	Six strain domains (i.e. family, peer, school/academic, economic, media and culture) based on general strain theory.	1,210,566 text documents, of which 350,314 (28.9%) contained terms related to children and adolescents.	Using text mining techniques, twenty-one strain topics were created by the authors and three other researchers grounded in the GST framework.
TENG AND VARATHAN (2023)	Ask.fm and upon comparison, AMiCA dataset	English	Cyberbullying (0) Non-Cyberbullying (1)	113,694 (5,375 Cyberbullying and 108,319 non-cyberbullying)	They used datasets made available by Van Hee et al., (2018). In this dataset, annotators with reliable expertise backgrounds participated, including four trained linguists in English and Dutch.

AL-HASHEDI ET AL. (2023)	Wikipedia and Twitter	English	CB and non-CB	25,000 tweets	Manually labeled by CrowdFlower
MURSHED ET AL. (2022)	Twitter	English	“0” non-cyber bullying or “1” cyberbullying	10000 tweets among which 6,508 (0.65%) are non-cyberbullying, and 3492 (0.35%) are cyberbullying tweets.	Tweets were labeled manually by three human annotators over a period of one and half months
DANIEL ET AL. (2023)	Twitter	Not specified	Offensive and Non-offensive	55788 tweets: Offensive (31353) and Non-offensive (24435)	EDL-TSGSO approach was used to classify different class labels.
OBAID, GUIRGUIS AND ELKAFFAS (2023)	Twitter	Not specified	Bullying and non-bullying	47,733 comments (39,748 were categorized as bullying and 7,985 as non-bullying)	A deep learning classifier was built for classifying a comment as cyberbullying or not.
AZEEZ ET AL. (2021)	Twitter	English	Cyberbullying or non-cyberbullying	350,000 tweets (Eight thousand tweets were selected at random from the cleaned dataset)	They were labelled with human coding with the assistance of three individuals familiar with twitter terminology, slang and parlance.
AL-MARGHILANI (2022)	Twitter	Not specified	Cyberbullying or non-cyberbullying	5076 instances in Cyberbullying class and 11,014 instances in Non-Cyberbullying class. Besides, dataset-2 has 20,620 instances in Cyberbullying class and 4163 instances in Non-Cyberbullying class.	Stacked auto encoder model is used as a classification model to allocate appropriate class labels of the data.
VAN HEE ET AL. (2018)	ASKfm	English and Dutch	Victim, bully, and two types of bystanders: bystander-defender and bystander-assistant	113,698 and 78,387 posts for English and Dutch, respectively	Manually annotated for cyberbullying according to a fine-grained annotation scheme.
BOZYIĞIT, UTKU AND NASIBOV (2021)	Twitter	Turkish	Non-harmful and cyberbullying contents	5000	Data annotation was done using a crowdsourcing web app by three computer science master's degree holders.
ROY AND MALI (2022)	Google and MMHS150K dataset	Not applicable	Bully and non-bully	3000 images containing 1458 bullying and 1542 not bullying images	With the help of three independent annotators
WU ET AL. (2023)	Twitter	Not specified	Annotated reviews were divided into six classes: religion, age, ethnicity, gender, non-cyberbullying, and other cyberbullying.	47,692	Not specified
BHARTI ET AL. (2022)	Twitter	Not specified	Cyberbullying and non-Cyberbullying	35,787 labeled tweets: 65% cyberbullying, 35% non-cyberbullying.	Not specified
PTASZYNSKI ET AL. (2019)	Data from unofficial school websites and forums	Japanese	Harmful and nonharmfull	1,490 harmful and 1,508 nonharmful entries.	The harmful and nonharmful sentences were manually labeled by Internet Patrol members according to official instructions included in the MEXT manual for dealing with cyberbullying (MEXT 2008).
ALDUAILAJ AND BELGHITH (2023)	Youtube and Twitter	Arabic	Bullying and non-bullying	30000 Arabic comments	The label was made depending on the most common and frequent bullying keywords in Arabic society, which were already collected manually from the frequent words in the posts.
PARUCHURI AND RAJESH (2023)	Twitter	English	Bullying and non-bullying	Bullying (523 testing, 5292 training) Non-bullying (1477 testing, 14.540 training) Total: 2000 testing, 19.832 training	based on Twitter annotations
ALDHYANI, AL-ADHAILEH AND ALSUBARI (2022)	(i) Wikipedia and (ii) Twitter	English	(i) Aggressive posts or non-aggressive posts (ii) religion, age, gender, ethnicity and non-cyberbullying tweets.	(i) 115,661 post samples, distributed as 101,082 aggressive posts and 14,782 non-aggressive posts (Wikipedia) (ii) 39,869 tweet samples, distributed across five types of online bullying classes	The authors used deep learning neural networks techniques to identify each social post as cyberbullying or not.
ZHANG, LI AND NAKAJIMA (2023)	Twitter	Japanese	Bullying and non-bullying	Dataset 1: 1766 bullying tweets 1766 non-bullying tweets Dataset 2: 1747 bullying tweets 1747non-bullying tweets	The authors used Yahoo! Japan crowdsourcing33 to label tweets. Each crowd worker is asked to label 10 tweets and each tweet is evaluated by three crowd workers.
LEPE-FAÚNDEZ ET AL. (2021)	Twitter	Spanish (Chilean, Mexican, and Chilean-Mexican corpora)	Aggressive and non-Aggressive	Chilean (2470), mexican (7332), chilean-mexican (9802)	Manually labelled.
RASOOL ET AL. (2023)	Twitter	Asian language	Non-cyberbullying ("0") or cyberbullying ("1") and offensive and non-offensive	31,353 offensive (56.2%) and 24,435 non-offensive (43.8%)	Three human annotators worked over 1.5 months to classify all the collected tweets.
ALGHAMDI ET AL. (2021)	Twitter	Arabic language	The first level classifies text into violent and non-violent. The second level classifies violent text into either cyberbullying or threatening.	3700 but after the cleaning process, the resulting dataset contains 2000	Tweets were annotated by two annotators following guidelines from psychology experts. A third annotator resolves disagreements for final labels.
DEWANI ET AL. (2023)	Twitter	Roman Urdu	Cyberbullying and non-cyberbullying	Not specified	The dataset is hand-annotated by linguistic experts and verified for data quality using Kohen's kappa.
RAJ ET AL. (2022)	Twitter	English, Hindi, and Hinglish	0-1 classifer	Not specified	Not specified

3.3 Cyberbullying applications

This subsection responds to the RQ3 and focuses exclusively on practical applications or prototypes that have been developed in recent years to detect cyberbullying, excluding work that is limited just to proposing models or comparing algorithms. We found 6 applications with one exclusively focused on children and adolescents through parental control, shown in Table 2:

Table 2: Applications for cyberbullying detection

AUTHORS	APPLICATION	TECHNIQUE
ZAMBRANO ET AL. (2023)	ParentIAl - Bullying (Parental control)	LDA Topic modeling
THUN, TEH AND CHENG (2022)	CyberAID (Mobile application)	Incorporates multiple features e.g. sentiment value, emoticon count, combination of personal pronouns and negative/profane words, among others, in a machine learning model using algorithms such as RF, DT, SVM and Naïve Bayes.
TENG AND VARATHAN (2023)	Cyberbullying Checker Application	Conventional ML and Transfer Learning
PTASZYNSKI ET AL. (2019)	Cyberbully Blocker (Mobile application)	Uses a combinatorial algorithm, similar to a brute force search algorithm, to automatically extract sophisticated sentence patterns and uses these patterns for textual classification of entries about cyberbullying.
LEPE-FAÚNDEZ ET AL. (2021)	A web app demonstrating model applicability for classifying tweets, evaluating implemented models, and gathering user feedback, while generating a database for future research.	5 approaches were used to create different models: Lexicon, TF_IDF_Lexicon, WE_Lexicon, WE_Lexicon_TF-IDF, and the Ensemble approach
RAJ ET AL. (2022)	A social media prototype for automatically detecting cyberbullying text on multilingual data (Hindi, English, Hinglish)	A deep neural network-based, using CNN-BiLSTM model

While many studies have demonstrated positive outcomes, the translation of these findings into practical applications accessible to society remains limited, as we can only access the application from Teng and Varathan (2023).

4 Final considerations

In summary, this study highlights the significant progress made in the field of cyberbullying detection using artificial intelligence, underscoring the advancements in research while acknowledging the need for investment in practical applications. By examining and synthesizing a wide range of studies with applications and algorithm comparisons, we were able to identify trends, gaps, and opportunities.

Despite studies such as Song et al. (2020), which used longitudinal information, we observed that in many proposed detection models, one issue was neglected: that cyberbullying is a systematic intimidation, that is, it occurs constantly or regularly. Analyzing just one tweet in isolation, for example, does not guarantee the detection of cyberbullying, but rather content that has the potential to constitute it. Additionally, the task of identifying some type of words represents only a subset of distinctive vocabulary and may not encompass all instances of cyberbullying. To avoid methodological inconsistencies, therefore, we suggest that detection models be considered as

“detection support models”, as, in fact, they can assist in detection, but are still not capable of effectively verifying the occurrence of systematic intimidation.

Furthermore, we emphasize that many datasets can be susceptible to bias, because of positive instances which can lead to unsatisfactory results if they are not properly trained on diverse and representative data. We consider bias as a challenge for future innovations in AI technologies, and not just in the context of cyberbullying. Daniel et al. (2023) also highlights the need to explore privacy-preserving AI techniques that can detect cyberbullying without compromising individuals’ personal information.

Roy and Mali (2022) also presented a gap in detection of cyberbullying through images, as in his proposal they considered only images but paid attention to the existence of combining the image with text in cyberbullying posts. This leads us to consider a major challenge for the future of research into cyberbullying detection, also considering other textual genres such as memes and even video content.

We also noted a lack of studies in Brazil focusing on cyberbullying detection and according to Eronen et al. (2022), the state-of-the-art results achieved for English are not representative globally. We therefore highlight the need for studies in the Brazilian context, considering the Portuguese language, its linguistic and regional variations, as well as the vocabulary used by young audiences, considering internet language, slang, gaming terms, among others. Furthermore, one must consider the most popular platforms in Brazil and the dynamics of interaction between young people in these communities. As future perspectives, therefore, we intend to develop a dataset for support detecting cyberbullying in Brazilian Portuguese in the context of children and adolescents. The identified gaps and opportunities can aid in proposing mitigating measures against cyberbullying in Brazilian primary and secondary school environments and support the development of a browser extension to support detect and notify the writing of potentially offensive messages.

A significant limitation of this systematic literature mapping is the exclusion of studies that consider Cyberbullying as a subtype of hate speech or cyber aggression, as only studies that use the term "cyberbullying" were considered. As a result, there may be a lack of representation of research that approaches the topic in different ways or contextualizes it within alternative conceptual frameworks. In addition, practical applications such as Abusniff (Talukder and Carbunar, 2018), BullyBlocker (Silva et al., 2018) and Rethink App (n.d., 2020) among others were not included in this study. Therefore, this study presents limitations due to the choice of search strings, the selection of databases and the period within which the publications were considered.

Finally, we believe that this review can help and encourage new research initiatives on cyberbullying and other problems that share some similar characteristics, such as hate speech, online racism, among others.

References

ALDHYANI, T. H. H.; AL-ADHAILEH, M. H.; ALSUBARI, S. N. Cyberbullying Identification System Based Deep Learning Algorithms. Electronics, v. 11, 2022.

ALDUAILAJ, A. M.; BELGHITH, A. Detecting Arabic Cyberbullying Tweets using Machine Learning. Mach. Learn. Knowl. Extr. 5, p.29-42, 2023.

ALGHAMDI, D.; AL-MOTERY, R.; ALMA'ABDI, R.; ALZAMZAMI, O. BABOUR, A. Automatic detection of cyberbullying and threatening in Saudi tweets using machine learning. International Journal of Advanced and Applied Sciences, v.8, n.10, 2021.

AL-HASHEDI, M.; SOON, L.; GOH, H.; LIM, A.H.L.; SIEW,E. Cyberbullying Detection Based on Emotion. IEEE Access, v. 11, 2023.

AL-MARGHILANI, A. Artificial Intelligence-Enabled Cyberbullying-Free Online Social Networks in Smart Cities. International Journal of Computational Intelligence Systems, v. 15, 2022.

AZEEZ, N.; IDIAKOSE, S.; ONYEMA, C.; VAN DER VYVER, C. Cyberbullying Detection in Social Networks: Artificial Intelligence Approach. Journal of Cyber Security and Mobility, v. 10, p. 745–774, 2021.

BALAKRISHNAN, V.; KHAN, S.; ARABNIA, H.R.; Improving cyberbullying detection using Twitter users' psychological features and machine learning. Computers & Security, v.90, 2020.

BALAKRISHNAN, V.; KHAN, S.; FERNANDEZ, T.; ARABNIA, H. Cyberbullying detection on twitter using Big Five and Dark Triad features. Personality and Individual Differences, v. 141, p. 252–257, 2019.

BAUMAN, S. (2015) Types of cyberbullying, Cyberbullying. https://doi.org/10.1002/9781119221685.ch4

BHARTI, S.; YADAV, A.K.; KUMAR, M.; YADAV, D. Cyberbullying detection from tweets using deep learning. Kybernetes, v. 51, n.9, p.2695-2711, 2022.

BOZYIĞIT, A.; UTKU, S.; NASIBOV, E. Cyberbullying detection: Utilizing social media features. Expert Systems With Applications, v. 179, 2021.

CETIC - CENTRO REGIONAL DE ESTUDOS PARA O DESENVOLVIMENTO DA SOCIEDADE DA INFORMAÇÃO SOB OS AUSPÍCIOS DA UNESCO. TIC Kids Online Brasil. Disponível em: <https://cetic.br/pt/pesquisa/kids-online/> Acesso em 22 abr. 2024

CHIA, Z. L.; PTASZYNSKI, M.; MASUI, F.; LELIWA, G.; WROCZYNSKI, M. Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. Information Processing and Management, v.58, 2021.

DANIEL, R.; MURTHY, S.; KUMARI, CH. D.V.P.; LYDIA, E. L.; ISHAK, M. K.; HADJOUNI, M.; MOSTAFA, S. M. Ensemble Learning With Tournament Selected Glowworm Swarm Optimization Algorithm for Cyberbullying Detection on Social Media. IEEE Access, v.11, 2023.

DATA REPORTAL (2023). Survey Digital 2023: Global Overview Report. Disponível em: <https://datareportal.com/reports/digital-2023-global-overview-report> Acesso em 10 de jul. 2023

DERMEVAL, D.; COELHO, J. A. P. de M.; BITTENCOURT, Ig I. Mapeamento Sistemático e Revisão Sistemática da Literatura em Informática na Educação. In: Metodologia de Pesquisa Científica em Informática na Educação: Abordagem Quantitativa. Porto Alegre: SBC, 2020. (Série Metodologia de Pesquisa em Informática na Educação, v. 2) Disponível em: <https://metodologia.ceie-br.org/livro-2> Acesso em: 04 fev. 2024

DEWANI, A.; MEMON, M.A.; BHATTI, S.; SULAIMAN, A.; HAMDI, M.; ALSHAHRANI, H.; ALGHAMDI, A.; SHAIKH, A. Detection of Cyberbullying Patterns in Low Resource Colloquial Roman Urdu Microtext using Natural Language Processing, Machine Learning, and Ensemble Techniques. Appl. Sci. v.13, 2023.

ERONEN, J.; PTASZYNSKI, M.; MASUI, F.; SMYWIŃSKI-POHL, A.; LELIWA, G.; WROCZYNSKI, M. Improving classifier training efficiency for automatic cyberbullying detection with Feature Density. Information Processing and Management, v. 58, 2021.

EVANS-LACKO, S.; TAKIZAWA, R.; BRINCLECOMBE, N.; KING, D.; KNAPP, M.; MAUGHAN, B. & ARSENEAULT, L. Childhood bullying victimization is associated with use of mental health services over five decades: a longitudinal nationally representative cohort study. Psychological Medicine, v.47, n1, p. 127-135, 2017.

GHOSH, A. & VEALE, T. Fracking sarcasm using neural network. In: Proceedings of NAACL-HLT 2016, Association for Computational Linguistics, 2017.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, 2007.

LEPE-FAÚNDEZ, M.; SEGURA-NAVARRETE, A.; VIDAL-CASTRO, C.; MARTÍNEZ-ARANEDA, C.; RUBIO-MAZANO, C. Detecting Aggressiveness in Tweets: A Hybrid Model for Detecting Cyberbullying in the Spanish Language. Appl. Sci. v. 11, 2021.

LÓPEZ-VIZCAÍNO, M.; NÓVOA, F.; CARNEIRO, V.; CACHEDA, F. Early detection of cyberbullying on social media networks. Future Generation Computer Systems v.118, p.219–229, 2021.

MEXT (2008). 'Netto-jō no ijime' ni kansuru taiō manyuaru jirei shū (gakkō, kyōin muke) ["Bullying on the Net" Manual for handling and collection of cases (for schools and teachers)] (in Japanese) In Ministry of education, culture, sports, science and technology (MEXT). Disponível em: <https://www.mext.go.jp/b_menu/houdou/20/11/08111701/001.pdf> Acesso em: 10 de jul. 2024

MURSHED, B.A.H.; ABAWAJY, J.; MALLAPPA, S.; SAIF, M. A.N.; AL-ARIKI, H.D.E. DEA-RNN: A Hybrid Deep Learning Approach for cyberbullying Detection in Twitter Social Media Platform. IEEE Access, v. 10, 2022.

OBAID, M. H.; GUIRGUIS, S. K.; ELKAFFAS, S. M. Cyberbullying Detection and Severity Determination Model. IEEE Access, v. 11, 2023.

PARUCHURI, V. L.; RAJESH, P. CyberNet: a hybrid deep CNN with N-gram feature selection for cyberbullying detection in online social networks. Evolutionary Intelligence v.16, p. 1935-1949, 2023.

PRICE, M. & DALGLEISH, J. Cyberbullying: experiences, impacts and coping strategies as described by Australian young people. Youth Studies Australia v.29, p. 51-59, 2010.

PTASZYNSKI, M.; LEMPA, P.; MASUI, F.; KIMURA, Y.; RZEPKA, R. Brute - Force Sentence Pattern Extortion from Harmful Messages for Cyberbullying Detection, v. 20, n.8, 2019.

RAJ, M.; SINGH, S.; SOLANKI, K.; SELVANAMBI, R. An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques. SN Computer Science, v. 3, 2022.

RASOOL, H. A.; ALDOLAIMY, F.; HASAN, F. F.; ALSALAMY, A. H.; SALEEM, M.; ALKHAYYAT, A. H.; SHARMA, M. Evolutionary Algorithm with Graph Neural Network Driven Cyberbullying Detection on Low Resource Asian Languages. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023.

RETHINK - STOPS CYBERBULLYING. Google Play App. ReThink. Disponível em: <https://play.google.com/store/apps/details?id=com.rethink.app.rethinkkeyboard&hl=en> Acesso em: 22 abr. 2024.

ROY, P. K.; MALI, F. U. Cyberbullying detection using deep transfer learning. Complex & Intelligent Systems v.8, p.5449–5467, 2022.

SAINJU, K. D.; MISHRA, N.; KUFFOUR, A.; YOUNG, L. Bullying discourse on Twitter: An examination of bully-related tweets using supervised machine learning. Computers in Human Behavior, v. 120, 2021.

SANGWAN, S. R,; BATHIA, MPS. Denigration Bullying Resolution using Wolf Search Optimized Online Reputation Rumour Detection. Procedia Computer Science, v. 173, p. 305–314, 2020.

SECURITY.ORG. Cyberbullying: Twenty Crucial Statics for 2024. [S.I] 6 jun 2024. Disponível em: <Cyberbullying: Twenty Crucial Statistics for 2024> Acesso em: 10 jul. 2024.

SILVA, Y.N, DEBORAH, L.H., RICH, C. BullyBlocker: Toward an Interdisciplinary Approach to Identify Cyberbullying. Social Network Analysis and Mining v.8, n.1, p.1–15, 2018.

SONG, J.; HAN, Y.; KIM, K.; SONG, T. M. Social big data of future signals for bullying in South Korea: Application of general strain theory. Telematics and Informatics, v. 54, 2020.

SONG, T.; SONG, J. Prediction of risk factors of cyberbullying-related words in Korea: Application of data mining using social big data. Telematics and Informatics, v. 58, 2021.

SWETHA, L. & RAVIKUMAR, S. A review on Fuzzy set Theory and its operations. International Journal of Advanced Research Trends in Engineering and Technology (IJARTET). Vol. 4, Special Issue 21, August 2017.

TALUKDER, S.; CARBUNAR, B. AbuSniff: Automatic Detection and Defenses Against Abusive Facebook Friends. Proceedings of the International AAAI Conference on Web and Social Media, v. 12, n.1, 2018.

TENG, T.H.; VARATHAN, K.D. Cyberbullying Detection in Social Networks: A comparison Between Machine Learning and Transfer Learning Approaches. IEEE Access, v. 11, 2023.

THUN, L. J.; TEH, P. L.; CHENG, C. CyberAid: Are your Children safe from cyberbullying? Journal of King Saud University, v,34, p.4099-4108, 2022.

UNESCO. School violence: Cyberbullying is a growing concern for safety and mental well-being. [S.I], 6 de nov. 2023. Disponível em: https://world-education-blog.org/2023/11/02/school-violence-cyberbullying-is-a-growing-concern-for-safety-and-mental-well-being/ Acesso em: 10 de jul. 2024

UNITED NATIONS. Cyberbullying a top concern on Safer Internet Day. [S.I]. 08 de fev. de 2022. Disponível em:

<https://news.un.org/en/story/2022/02/1111512#:~:text=They%20were%20asked%20to%20rank,and%2026%20per%20cent%20respectively> Acesso em: 04 de fev. 2024

VAN HEE, C. JACOBS, G.; EMMERY, C.; DESMET, B.; LEFEVER, E.; VERHOEVEN, B.; DE PAUW, G.; DAELEMANS, W.; HOSTE, V. Automatic detection of cyberbullying in social media text. PLoS ONE, v. 13, 2018.

WU, F.; GAO, B.; PAN, X.; SU, Z.; JI, Y.; LIU, S.; LIU, Z. FACapsnet: A fusion capsule network with congruent attention for cyberbullying detection. Neurocomputing, v.542, 2023.

WORLD HEALTH ORGANIZATION (WHO). One in six school-aged children experiences cyberbullying, finds new WHO/Europe study. Disponível em: <https://www.who.int/europe/news/item/27-03-2024-one-in-six-school-aged-children-experiences-cyberbullying--finds-new-who-europe-study#:~:text=15%25%20of%20adolescents%20(around%201,%25%20to%2016%25%20for%20girls.> Acesso em 22 abr. 2024

YUVARAJ, N.; CHANG, V.; GOBINATHAN, B.; PINAGAPANI, A.; KANNAN, S.; DHIMAN, G. RAJAN, A. Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. Computers and Electrical Engineering v.92, 2021.

ZAMBRANO, P.; TORRES, J.; TELLO-OQUENDO, L.; YÁNEZ, Á.; VELÁSQUEZ, L. On the modeling of cyber-attacks associated with social engineering: A parental control prototype. Journal of Information Security and Applications, v.75, 2023.

ZHANG, J.; LI, L.; NAKAJIMA, S. Constructing Japanese Bullying Expression Dictionary for Automated Cyberbullying Detection on Twitter. Vietnam Journal of Computer Science, v. 10, n.2, p. 135-158, 2023.

Data de submissão: 10/05/2024 | Data de revisão: 20/06/2024 | Data de aceite: 28/06/2024 | Data de publicação: 31/07/2024