

# Cybersecurity Approaches to Manage Cyberbullying - A Systematic Review

## Introduction:

As of 2020 there are estimated to be more than four billion internet users across the globe, highlighting an ever increasing level of embeddedness throughout society (Bashir Shaikh, Rehman and Amin, 2020). With so many users spread across a diverse range of ages, socioeconomic backgrounds and demographics it should come as no surprise that this high level of connectedness and internet use has led to increased levels of internet abuse, giving rise to the phenomenon of cyberbullying.

Cyberbullying, broadly defined, embodies a range of abuse or harassment carried out by one or many individuals via a digital medium such as SMS texting, social media, comment sections or web forums (Yokotani and Takano, 2022). With the scope of web-platforms and the relative anonymity these platforms afford, cyberbullying represents a current and growing risk to the online community.

This literature review aims to assess the cybersecurity field's current approach to the detection of cyberbullying using machine learning as well as some of the challenges researchers face addressing a problem redelent across many cultures, languages and ages.

## Methodology:

The process of performing a systematic literature review (SLR) encompasses three phases:

1. Planning
2. Conducting
3. Reporting (Drummond and Machado, 2021)

### *Planning Phase:*

The planning phase was largely carried out by using the open source research planner Parsifal to aid in the formulation of the core research questions as well as to generate and test a suitable search string to better target more relevant literature from academic sources including but not limited to *ResearchGate*, *IEEE* and *Scopus*.

To better define the research questions, PICOC (Population, Intervention, Comparison, Outcome, Context) criteria was outlined as follows:

- **Population:** Cybersecurity, cyber security
- **Intervention:** ai, tools, artificial intelligence, machine learning
- **Comparison:**
- **Outcome:** cyberbullying
- **Context:** social media, education

This framed our two primary research questions used to critically review and analyze a selected body of literature:

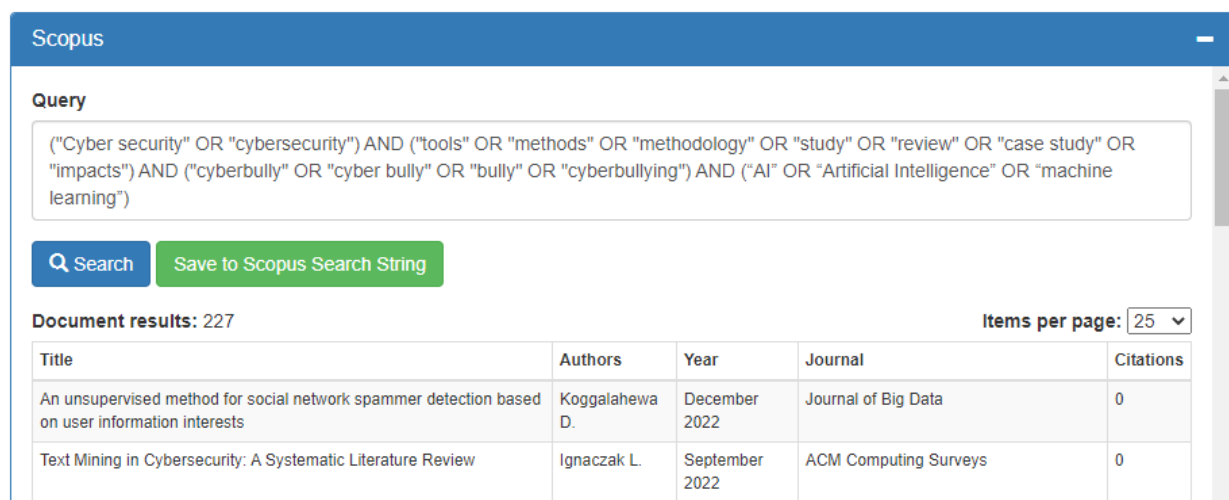
*RQ1: What is the efficacy of applying machine learning principles to the detection of cyberbullying across social media communities?*

*RQ2: How does the cybersecurity research community address cyberbullying as a universal phenomenon, distinguished by a diversity of languages, cultures and political ideologies?*

Literature selection criteria for this review only included academic sources that were either published in a journal or as part of a conference publication. Gray sources were excluded as well as publications dating further back than 2015. All cited sources were published in English.

Parsifal was used to define the PICOC criteria and research questions that encompass this review. Parsifal was also used to generate a search string that could be easily tested on the Scopus database. While this platform is also capable of querying IEEE and Research Gate databases, the feature was not functioning at the time of writing this review.

- ("Cyber security" OR "cybersecurity") AND ("tools" OR "methods" OR "methodology" OR "study" OR "review" OR "case study" OR "impacts") AND ("cyberbully" OR "cyber bully" OR "bully" OR "cyberbullying") AND ("AI" OR "Artificial Intelligence" OR "machine learning")



The screenshot shows the Scopus search interface. At the top, the 'Scopus' logo is visible. Below it, the 'Query' section contains the search string: ("Cyber security" OR "cybersecurity") AND ("tools" OR "methods" OR "methodology" OR "study" OR "review" OR "case study" OR "impacts") AND ("cyberbully" OR "cyber bully" OR "bully" OR "cyberbullying") AND ("AI" OR "Artificial Intelligence" OR "machine learning"). Below the query, there are two buttons: 'Search' and 'Save to Scopus Search String'. The 'Document results: 227' section shows a table with 5 columns: Title, Authors, Year, Journal, and Citations. The table lists two results: 'An unsupervised method for social network spammer detection based on user information interests' by Koggalahewa D. (December 2022, Journal of Big Data, 0 citations) and 'Text Mining in Cybersecurity: A Systematic Literature Review' by Ignaczak L. (September 2022, ACM Computing Surveys, 0 citations).

Title	Authors	Year	Journal	Citations
An unsupervised method for social network spammer detection based on user information interests	Koggalahewa D.	December 2022	Journal of Big Data	0
Text Mining in Cybersecurity: A Systematic Literature Review	Ignaczak L.	September 2022	ACM Computing Surveys	0

Figure 1. Output from Parsif.al search string query

Satisfied with the output of the query, the search string was used manually to find relevant literature across our chosen source databases, as well as the University of Essex library. In some circumstances the string was altered to better define the specific database's results and limit the “noise” of some irrelevant results. To adhere to the time limitations of this literature review, twenty-one papers were selected for review. Source literature selection emphasized papers that highlighted machine learning and alternative methodologies being used to address cyberbullying.

#### *Conducting Phase:*

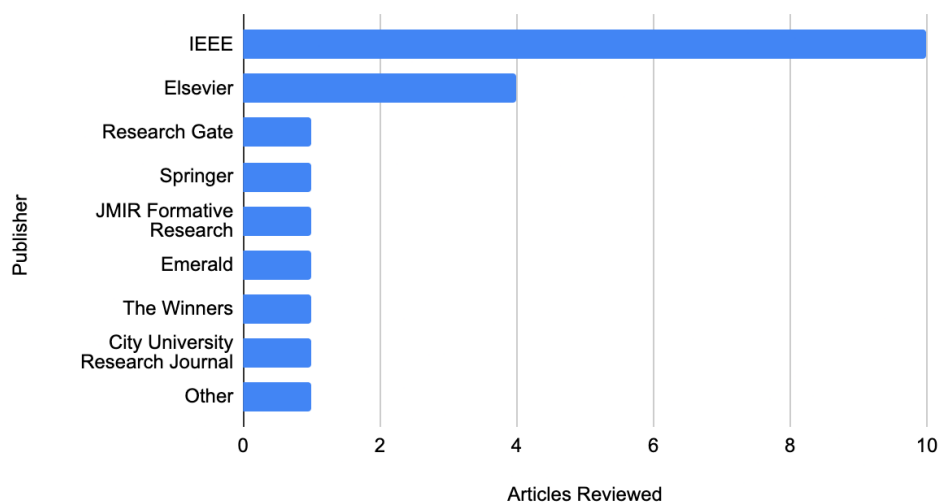
Analysis and data extraction was performed by manually reviewing the selected material and collecting data through review. Notes were taken to flesh out trends and themes that aid in answering the research questions.

### **Results and Discussion:**

#### ***Review of Source Material***

The chart below shows the breadth of sourced material for review. While a range of publication sources were used, IEEE garnered the most results accounting for 47% of the referenced material.

Distribution of Source Publisher



*Figure 2. Distribution of Articles by Publication*

Further to publication source, the publications used dated back no further than 2015, as seen in **Figure 3** where articles span 2015 to 2022. The majority of sources used were published in 2018 (23.8%) and 2021 (33.3%).

Distribution of Articles by Year

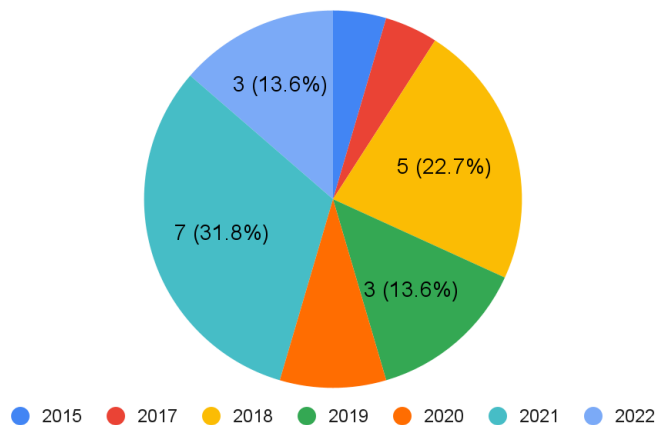


Figure 3. Distribution of Articles by Year

### Results of Research Questions:

#### ***RQ1: What is the efficacy of applying machine learning principles to the detection of cyberbullying across social media communities?***

Of the twenty-one articles reviewed, 18 focused on machine or deep learning approaches to detecting cyberbullying across social media and forum databases. A resurfacing challenge of using machine learning in cyberbullying detection is the availability of these datasets to train the researcher's algorithm; a fact exacerbated when the language does not use the Roman alphabet such as Arabic, Chinese (Li, 2019) or Japanese (Ferdous Khan, Inoue and Sakamura, 2021).

A number of the studies achieved promising results, especially when employing the use of Naive-Bayes (NB) classifiers and support-vector machine (SVM) algorithms to analyze datasets. Three of the fifteen machine learning oriented papers employed the use of NB and SVM. (Kumar, 2021) found the accuracy for NB to surpass 96% and SVM to be 97% accurate when applied to MySpace, SlashDot and Kongregate datasets. While the algorithms were found accurate on the above datasets, others proved less so. For example in Kumar's research, these same algorithms were found to be 60% and 67% accurate in detecting cyberharassment when applied to Youtube and Formspring datasets respectively (Kumar, 2021).

Studies often employed bespoke algorithms applying to controlled or specific circumstances. Examples of which can be seen in (Toapanta Toapanta, Alfredo Espinoza Carpio and Mafla Gallegos, 2020) where the targeted form of cyberbullying only included profanity as a form of harassment, acknowledging the difficulty that

context provides where the model failed to identify insults in jest relative to true harassment. Another challenge in assessing incidents of cyberbullying using automated methodology is the requirement to massage responses by removing images, alphabetical casing and emojis, which further reduces the context of responses. An example of this being if you commented on a friend's football preferences by insulting them, while including wink or tongue emoji. Machine learning generally requires this emoji or image context to be stripped from the dataset leading to false harassment flags or an exorbitant amount of manual analysis in an attempt to understand situational context (Gorro et al., 2018).

A recurring theme among researchers applying machine/deep learning to cyberbullying detection is the need for larger datasets to train algorithms. This would lead to more accurate results and allow for a more scalable range of detectable offenses. Seven of the eighteen (39%) machine learning articles called for future analysis using larger datasets to improve results. Alotaibi, Alotaibi and Razaque's 2021 study believed they could increase model accuracy as well as the scope of detection capabilities across different types of offenses carried out in different languages; a statement echoed throughout the material analyzed for this review (Alotaibi, Alotaibi and Razaque, 2021).

The above findings suggest that while there is significant promise in applying machine learning algorithms to detect cyberbullying across web platforms it is not without its challenges. The variability of success across different social media datasets as well as the challenges faced where there is a shortage of available data to assess non-English platforms meant the models had to be significantly tailored and required considerable manual intervention to be viable.

***RQ2: How does the cybersecurity research community address cyberbullying as a universal phenomenon, distinguished by a diversity of languages, cultures and political differences?***

In Bashir Shaikh, Rehman and Amin's 2020 paper, *A Systematic Literature Review to Identify the Factors Impelling University Students Towards Cyberbullying* the authors found that a wide range of environmental and societal factors played a contributing role in those inclined to perpetuate cyberbullying and those who are recipients of

cyberbullying. One factor that makes cyberbullying increasingly challenging to combat is the relative anonymity afforded by online communities and the propensity for bullies to be emboldened by the actions of other cyberbullies online. This quality appears to be compounded in discussions related to socio-political topics where trolling or harassment by one can lead to a mobbing of individuals with similar beliefs (Bashir Shaikh, Rehman

and Amin, 2020). This increased polarization of political beliefs led to a stark increase in cyberbullying during the start of the Covid-19 pandemic in Ecuador where increases in web use, socio-economic strain and increasingly isolated communities took their frustrations to internet communities (Toapanta Toapanta, Alfredo Espinoza Carpio and Mafla Gallegos, 2020).

This creates both a challenge of how to detect and address harassment that is as diverse and fluid as the day's headlines. This Ecuadorian study suggested housing data on the blockchain for secure storage and analysis to study the impacts of cyberbullying and to propose future work to build a means to identify cyberbullying as it occurs during potential future anthropogenic disasters.

In contrast to increased cyberbullying in Latin America, the culture of politeness in Japan has led to the proposal of an app or bot that allows users to run comments through an artificial intelligence (AI) engine that is capable of analyzing text for accidental rudeness or aggression before posting. Cultures have differing approaches to managing online delinquency and while a Japanese study proposing a rudeness detecting bot may seem silly, it is more logical within the framework of Japanese culture and legislation. Cyberbullying is taken very seriously and is considered a human rights issue. Yahoo Japan already uses AI to flag and remove harmful comments, so it is logical that there is interest in developing a platform to prevent hurtful comments before they are even posted (Ferdous Khan, Inoue and Sakamura, 2021).

In a number of countries the proposed method of addressing cyberbullying is simply to begin educating students during formative years of development in an attempt to dissuade youth from being future offenders. This is in part due to the challenge of obtaining data and resources in countries where spoken language makes analyzable data scarce (Setiawan et al., 2020). In a Korean study, there was consideration for mandatory ethics classes in cyber-social principles for teens. Researchers also acknowledged that those with increased exposure to web platforms increase their threat landscape, laying the burden of risk on those who choose to be more active online (Choi, Cho and Lee, 2019). Likewise, a Turkish study proposed manual education that included educating both student and teacher as surveys found educators had a limited awareness of cyberbullying, but had a rudimentary understanding of the impacts seen in those experiencing online bullying. This study proposed a range of solutions from teacher and student education, government policy, cyberbullying public service ads and web filtering throughout schools (Sezer, Yilmaz and Karaoglan Yilmaz, 2015).

## **Future Work:**

A recurring theme throughout this literature review has been the regionality of cybersecurity research in both technical and cyber-social approaches and outcomes. An under-researched area is the impact the internet and subsequent exposure to cyberbullying has had on developing communities or countries where usage and exposure may be different from those who have lived with online communities for decades. For those newer entrants to online communities, are the negative effects of cyberbullying more pronounced and if so, what methods of detection and deterrence offer the best solution. Is it an automated technical approach using machine learning and AI or is education and awareness the best approach to combating cyberbullying?

## **Conclusion:**

This literature review aimed to collect academically relevant literature from journals and conference papers in an attempt to analyze current trends in cyberbullying research. In order to answer our two research questions, journal and conference articles were selected based on their relevance to machine learning applications in cyberbullying detection as well as different geographic and sociological approaches to the study of cyberbullying.

Analysis on the body of literature found that there was promise in applying machine learning algorithms to detect cyberbullying across datasets, but considerable challenges still hamper the viability of some of these approaches. Algorithms found to be effective on some datasets were much less accurate on others, while manual data massaging was also required to strip images and emojis from data; further muddying the context of comments and leading to falsely flagged incidents of harassment or bullying.

Language, culture and the geopolitical climate were also contributing factors to both the detection, occurrence and prevention of cyberbullying. Languages that don't use alphabet characters such as Chinese, Japanese or Arabic had difficulty obtaining analyzable datasets relevant to their locales. Furthermore, the context and regional differences meant that combating cyberbullying became a regional issue; as seen in Ecuador, Turkey and Japan.

## References:

- Ali, W.N.H.W., Mohd, M. and Fauzi, F. (2021). Cyberbullying Predictive Model: Implementation of Machine Learning Approach. In: *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*. 2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP). Available at: <https://ieeexplore.ieee.org/document/9497932> [Accessed 25 Feb. 2022].
- Alotaibi, B., Alotaibi, M. and Razaque, A. (2021). *A Multichannel Deep Learning Framework for Cyberbullying Detection on Social Media*. ResearchGate. Available at: [https://www.researchgate.net/publication/355845224\\_A\\_Multichannel\\_Deep\\_Learning\\_Framework\\_for\\_Cyberbullying\\_Detection\\_on\\_Social\\_Media](https://www.researchgate.net/publication/355845224_A_Multichannel_Deep_Learning_Framework_for_Cyberbullying_Detection_on_Social_Media) [Accessed 11 Mar. 2022].
- Bashir Shaikh, F., Rehman, M. and Amin, A. (2020). Cyberbullying: A Systematic Literature Review to Identify the Factors Impelling University Students Towards Cyberbullying. *IEEE Access*, 8, pp.148031–148051. Available at: <https://ieeexplore.ieee.org/document/9163353> [Accessed 25 Feb. 2022].
- Drummond, B.M. and Machado, R.C.S. (2021). Cyber Security Risk Management for Ports - a Systematic Literature Review. In: *2021 International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea)*. 2021 International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea). IEEE. Available at: <https://ieeexplore.ieee.org/document/9611569> [Accessed 6 Mar. 2022].
- Ferdous Khan, M.F., Inoue, R. and Sakamura, K. (2021). Development of a Bot using Sentiment Dictionary and Machine Learning for Proactive Detection of Slanderous Japanese Posts on Social Media. In: *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*. 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE). IEEE, pp.559–563. Available at: <https://ieeexplore.ieee.org/document/9621837> [Accessed 27 Feb. 2022].
- Gorro, K.D., Sabellano, M.J.G., Gorro, K., Maderazo, C. and Capao, K. (2018). Classification of Cyberbullying in Facebook Using Selenium and SVM. *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, pp.183–186.
- Kumar, R. (2021). Detection of Cyberbullying Using Machine Learning. *Turkish Journal of Computer and Mathematics (TURCOMAT)*, 12(9), pp.656–661. Available at: <https://turcomat.org/index.php/turkbilmat/article/download/3131/2693> [Accessed 27 Feb. 2022].



Li, W. (2019). A Design Approach for Automated Prevention of Cyberbullying Using Language Features on Social Media. In: *IEEE Xplore*. [online] 5th International Conference on Information Management (ICIM). 2019 5th International Conference on Information Management (ICIM), pp.87–91. Available at: <https://ieeexplore.ieee.org/document/8714683> [Accessed 25 Feb. 2022].

Setiawan, W.V., Fitrisna, V.E., Michellianouva, F. and Mayliza, C.S. (2020). Cyberbullying Phenomenon of High School Students: An Exploratory Study in West Kalimantan, Indonesia. *The Winners*, 21(1), p.15.

Sezer, B., Yilmaz, R. and Karaoglan Yilmaz, F.G. (2015). Cyber bullying and teachers' awareness. *Internet Research*, 25(4), pp.674–687. Available at: [www.emeraldinsight.com/1066-2243.html](http://www.emeraldinsight.com/1066-2243.html) [Accessed 25 Feb. 2022].

Toapanta Toapanta, S.M., Alfredo Espinoza Carpio, J. and Mafla Gallegos, L.E. (2020). *An Approach to Cybersecurity, Cyberbullying in Social Networks and Information Security in Public Organizations during a Pandemic: Study case COVID-19 Ecuador*. IEEE Xplore. Available at: <https://ieeexplore.ieee.org/abstract/document/9240375> [Accessed 25 Feb. 2022].

Yokotani, K. and Takano, M. (2022). Predicting Cyber Offenders and Victims and Their Offense and Damage Time from Routine Chat Times and Online Social Network Activities. *Computers in Human Behavior*, 128(128), pp.1–15. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0747563221004222?via%3Dihub> [Accessed 27 Feb. 2022].

## **Bibliography:**

Ademiluyi, A., Li, C. and Park, A. (2022). Implications and Preventions of Cyberbullying and Social Exclusion in Social Media: Systematic Review. *JMIR Formative Research*, 6(1), p.e30286. Available at: <https://formative.jmir.org/2022/1/e30286> [Accessed 4 Feb. 2022].

Crepax, T., Muntés-Mulero, V., Martinez, J. and Ruiz, A. (2022). Information Technologies Exposing Children to Privacy risks: Domains and children-specific Technical Controls. *Computer Standards & Interfaces*, 82, p.103624. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0920548922000058> [Accessed 14 Mar. 2022].

Koggalahewa, D., Xu, Y. and Foo, E. (2022). An unsupervised method for social network spammer detection based on user information interests. *Journal of Big Data*,

9(1). Available at:

<https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-021-00552-5.pdf>  
[Accessed 14 Mar. 2022].

Koutsos, T.M., Menexes, G.C. and Dordas, C.A. (2019). An efficient framework for conducting systematic literature reviews in agricultural sciences. *Science of The Total Environment*, 682, pp.106–117. Available at: <https://pubmed.ncbi.nlm.nih.gov/31108265/>  
[Accessed 9 Feb. 2022].

Kramer, L. (2021). *How to Write a Stellar Literature Review*. How to Write a Stellar Literature Review | Grammarly Blog. Available at:  
[https://www.grammarly.com/blog/literature-review/?gclid=Cj0KCQiAxoiQBhCRARIsAPsvo-zQ-XXJRWOMtnRMG0y6aeYViVAa04ii\\_fNQkb4q8C\\_EdFswk8Y9DalaAsYIEALw\\_wcB&gclid=aw.ds](https://www.grammarly.com/blog/literature-review/?gclid=Cj0KCQiAxoiQBhCRARIsAPsvo-zQ-XXJRWOMtnRMG0y6aeYViVAa04ii_fNQkb4q8C_EdFswk8Y9DalaAsYIEALw_wcB&gclid=aw.ds) [Accessed 16 Feb. 2022].

Murnion, S., Buchanan, W.J., Smales, A. and Russell, G. (2018). Machine learning and semantic analysis of in-game chat for cyberbullying. *Computers & Security*, 76, pp.197–213. Available at:  
<https://www.sciencedirect.com/science/article/pii/S0167404818301597> [Accessed 2 Mar. 2022].

Qaisar, S. (2020). How deep was the cyberbullying episode? The mediating effect of social vulnerability between perceived cyberbullying severity and employee's behavioral outcomes at workplace: the moderating role of self-regulation. *CITY UNIVERSITY RESEARCH JOURNAL*, 10(4). Available at:  
<https://cusitjournals.com/index.php/CURJ/article/view/304> [Accessed 14 Mar. 2022].

Yokotani, K. and Takano, M. (2022). Predicting cyber offenders and victims and their offense and damage time from routine chat times and online social network activities. *Computers in Human Behavior*, [online] 128, p.107099. Available at:  
<https://www.sciencedirect.com/science/article/abs/pii/S0747563221004222> [Accessed 27 Feb. 2022].