**Tools and Techniques to reduce cyberbullying: A review of the Literature**

# **Introduction**

Cyberbullying refers to the use of technology or the internet to harass or intimidate another person (Giumetti & Kowalski, 2022). This can include sharing rumours, spreading abusive images on social media and sending mean or hurtful messages via text or email (Giumetti & Kowalski, 2022). Statista ranked cyberbullying among the leading online abuse in 2022, with 51% of respondents reporting being victims (Dixon, 2022). A survey by Comparitech found that offensive name-calling (31%) and purposeful embarrassment (26%) as the leading forms of cyberbullying (Cook, 2022). The review critically analyses and examines credible sources exploring cybersecurity tools and techniques to reduce cyberbullying. The review will also critically evaluate existing literature, research methodologies, commonalities and differences in the findings, research gaps, and limitations. The audience of this review is cyberbullying victims in the UK and globally.

One in ten people in the UK has experienced some form of cyberbullying – a figure rising annually (Giumetti & Kowalski, 2022). This figure is staggering thus it is significant to analyse models and techniques currently used to detect and reduce cyberbullying. It could provide insight by comparing the accuracy of various machine learning models in detecting cyberbullying on social media.

The Systematic Literature Review (SLR) framework was used to synthesise the literature in this review. SLR focuses on identifying and synthesising knowledge from scholarly and peer-reviewed sources derived from reputable sources. SLR focuses on relevant and quality sources related to the research topic (Reddy et al., 2022). The perspective is that combining various machine learning models enhances precision in detecting and reducing cyberbullying (Reddy et al., 2022). Sources considered in the last five years include academic papers, and journals, found on ACM Digital Library and Google Scholar. The findings, strengths and limitations, and discrepancies are discussed.

**2.0 Findings**

Cyberbullying has been identified as a cybersecurity risk in the context of online security, especially on social media. Despite the lack of substantive evidence in current literature linking cyberbullying to significant cyberattacks, deploying effective cybersecurity awareness programs is necessary (Khader, Karam, and Fares, 2021). Private and public organisations have deployed unique cybersecurity tools and techniques to manage cyberbullying effectively. As an example, in some schools in UAE, cyberbullying is integrated as a component of cybersecurity awareness programs for students (Khader, Karam, and Fares, 2021).

A case study by Parkin and Chua (2020) shows how social media controls can effectively reduce cybercrimes and online abuses. Cybersecurity controls' implications on legitimate users are often overlooked (Parkin & Chua, 2020). Key factors contributing to abusive online behaviour include pro-victim attitudes and the context of exchange. Anonymity, the distance between the users, and internet scalability also determine online abusive behaviours (Parkin & Chua, 2020). The scalability of the internet enables multiple users to interact and engage in bullying activities anonymously. The findings of Parkin and Chua (2020) are consistent with Faucher, Cassidy, and Jackson's (2020) results. However, the research methodologies used are different. Parkin & Chua (2020) used a case study, while Faucher, Cassidy, and Jackson (2020) used a qualitative thematic analysis of students' comments from surveys and focus groups. Parkin and Chua (2020) and Faucher, Cassidy, and Jackson (2020) agree that anonymous online interactions contribute to increased cyberbullying.

The SCENE framework discussed by Parkin and Chua (2020) leverages existing risk management tactics and behavioural change to reduce cybercrime. Future research should investigate the notion of social-technical precision in cyber risk management and cybersecurity (Parkin & Chua, 2020). Relevant stakeholders in a real-world environment, such as social media, should be involved. The findings of Parkin and Chua (2020) are consistent with the results of a study by Cheng et al. (2020). Cheng et al. (2020) experimented with cyberbullying detection using unsupervised learning models and data sets crawled from Instagram and Vine.

In a similar case study, Chua et al. (2019) applied the unintended harms framework to address cyberbullying. The findings show that cyberbullying victimisation correlates with various negative consequences. The prevalence of cyberbullying is linked to the increased connectivity of users through the Internet and social media platforms. For instance, victims of cyberbullying are more likely to report problematic behaviours offline (Chua et al., 2019). Chua et al. (2019) findings are inconsistent with the results of Cheng et al. (2020), which emphasise the effectiveness of unsupervised learning methods for detecting and addressing cyberbullying.

The training for teenagers establishes fundamental knowledge of cyberbullying and helpful online behaviour (Chua et al., 2019). For the parents and teachers, the training focuses on appropriate prevention and response mechanisms (Chua et al., 2019). Privacy control and management emphasise the recommended behaviour to enhance privacy online. Teenagers are encouraged to maximise privacy settings and controls on their devices and avoid sharing personal information to prevent cyberbullying at school and home. The results of Chua et al. (2019) are consistent with the findings of a systematic literature review by Quayyum, Cruzes, and Jaccheri (2021), showing that creating awareness is an effective strategy to reduce cyberbullying.

According to Talpur and O'Sullivan (2020), millions of active users communicate daily on Twitter. Users with Twitter accounts created between 2-5 years are the most vulnerable to cyberbullying, including sexual content (Talpur and O'Sullivan, 2020, p.16). Alotaibi, Alotaibi, and Razaque (2021) proposed an automatic cyberbullying method using a consolidated deep-learning mode to detect aggressive behaviour. The technique leverages multi-channel deplaning based on three models, the Bidirectional Gated Recurrent Units (BiGRU), Convolutional Neural Network (CNN), and Transformer Block (Alotaibi, Alotaibi, and Razaque, 2021). The technique classifies Twitter comments into aggressive and non-aggressive categories. The performance of the proposed method was evaluated using three popular hate speech data sets, achieving 88% accuracy.

Shape

Description automatically generated with medium confidence

**Figure 1: Comparison of the Different Methods** (Alotaibi, Alotaibi, and Razaque, 2021, p.12)

Figure 1 shows how the proposed method compares with the performance of existing methods. The method proposed by Alotaibi, Alotaibi, and Razaque (2021) has an accuracy of 87.99%, a better performance than BiGRU and CNN with 87.43% and 87.28%, respectively. The method for detecting cyberbullying on Twitter was also evaluated using precision, recall, F1-score, and confusion matrix (Alotaibi, Alotaibi, and Razaque, 2021). The technique achieved 87%, 85%, and 86% precision, recall, and F-score, respectively.

Comparably, a deep learning detection system developed by Gamback et al. (2017) uses CNN to classify tweets into racism, sexism, and non-offensive categories. The method uniquely detects offensive and non-offensive tweets. The solution by Alotaibi, Alotaibi, and Razaque (2021), combines the three techniques to classify offensive and non-offensive Twitter comments. As a result, offensive tweets containing content categorised as bullying can be effectively blocked. Consequently, combining the different deep learning models enhances accuracy in detecting and preventing cyberbullying (Gamback et al. 2017).

Chart, bar chart

Description automatically generated**Figure 2: Evaluation Metrics and Confusion Matrix** (Alotaibi, Alotaibi, and Razaque, 2021)

Figure 2 shows the evaluation metrics and the confusion matrix for the new cyberbullying detection method. The precision-recall F1 shows the accuracy scores in detecting offensive instances on Twitter (Alotaibi, Alotaibi, and Razaque, 2021).

# **3.0 Strengths and Limitations**

The proposed tools and techniques to address cyberbullying have unique strengths and limitations. A cyberbullying detection technique proposed by Alotaibi, Alotaibi, and Razaque (2021) combines three data sets from Davidson et al. (2017), Waseem and Hovy (2016), and Kaggle (2018). Combined datasets are comprehensive and recommended for accurate modelling. The results can be generalised to detect cyberbullying on Twitter and social media platforms effectively. Alotaibi, Alotaibi, and Razaque (2021) recommend that future research use a larger dataset to enhance the method's performance. Deep learning algorithms and models could be effectively and accurately trained with large datasets.

Additionally, combining quantitative and qualitative analysis techniques is a vital strength of a study by Cheng et al. (2020). Cheng et al. (2020) introduced an innovative model for detecting cyberbullying using unsupervised learning models combining a representation learning network and multitask learning network. The machine learning models leverage social media's hierarchical structure to detect cyberbullying. Combining qualitative and quantitative analysis provides a comprehensive outlook of how the proposed model competitively performs compared to supervised models.

Previous studies experimented with how various machine-learning algorithms can be used to detect cyberbullying. Alsubait and Alfageh's (2021) study used machine learning algorithms to detect cyberbullying in Arabic YouTube comments. The machine learning models compared are Multinomial Naive Bayes (MNB), Linear Regression (LR), and Complement Naive Bayes (CNB). Alsubait and Alfageh (2021) investigated two specific feature extraction methods, Count Vectorizer and Tfidf Vectorizer. The result shows that a combination of CV and LR models outshines MNB and CNB in performance.

Remarkably, Alsubait and Alfageh (2021) presented the details of the experiments compared to the performance of the three machine learning methods. The source of the dataset used is precisely indicated. The dataset contained more than 15,000 YouTube videos from Arabic users. A substantial explanation for using YouTube comments over Tweets was provided. However, the comments were collected between 2015 and 2017 (Alsubait and Alfageh, 2021). Collecting recent datasets from YouTube will enhance the accuracy of future studies to detect cyberbullying. Alsubait and Alfageh (2021) emphasise that combining various machine learning methods improves performance and accuracy for detecting cyberbullying.

Similarly, Cheng et al. (2021) proposed an innovative model for detecting cyberbullying on social media sessions using a hierarchical attention network. A social media session is hierarchical and multimodal. A comment consists of a sequence of words; each session has timestamps and key social content. The approach proposed constructs social media sessions in a bottom-up design to exclusively model its structure hierarchically. The model accounts for the order of words and comments and the attention level for each comment Cheng et al. (2021).

Diagram

Description automatically generated

**Figure 3: Hierarchical Structure for social media** (Cheng et al., 2021, p.8)

Figure 3 illustrates the hierarchical structure and attention methods for detecting cyberbullying in social media sessions.

# **4.0 Discrepancies**

Several discrepancies have been identified in the literature reviewed. For instance, Cheng et al. (2021) highlight three critical inconsistencies with the proposed technique for detecting cyberbullying. Integrating temporal information, such as the comments' timestamps, into cyberbullying detection is complex (Cheng et al., 2021). Achieving high accuracy in cyberbullying detection requires the incorporation of various dynamics of a social media session.

The second challenge is the scarcity of comment-level labels. Cheng et al. (2021) note that most of the existing datasets publicly do not include comment-level labels, whether a comment is a cyberbully. The findings by Cheng et al. (2021) are consistent with the results of a study by Singh and Hofenbitzer (2019). Out of the 4,865 messages from the Twitter dataset, only 93 (nearly 2%) were labelled bullying messages (Singh and Hofenbitzer, 2019, p.558). Therefore, this is a discrepancy that future research on cyberbullying detection should explore.

Ultimately, the multimodal nature of social media sessions complicates accurately detecting cyberbullying in social media. According to Cheng et al. (2021), social media data is a sea of noisy, informal, and short information. Hence, future research should focus more on developing adaptative and robust techniques and models to detect cyberbullying on social media.

Finally, most of the studies provided brief recommendations for future research. For instance, the suggestions by Singh and Hofenbitzer (2019) are unclear to researchers interested in improving the study. Singh and Hofenbitzer (2019) only offer a single stamen highlighting the need to consider more extensive networks, diverse operationalisation, and debiasing approaches. Providing articulated recommendations will enable future researchers to expound on the study effectively. The research discrepancy was also identified in a survey by Kaluarachchi, Warren, and Jiang (2020). Kaluarachchi, Warren, and Jiang (2020) highlight how existing literature prioritises education and skill development to deal with cyberbullying, especially among teens.

# **5.0 Conclusion**

Cyberbullying is a cybersecurity risk and threat prevalent online, especially on social media platforms such as Twitter, Facebook, and Instagram (Giumetti & Kowalski, 2022). Various frameworks and models have been proposed to detect, prevent, and reduce cyberbullying. Most of the previous sources experimented with how different machine-learning models can be used to detect and prevent cyberbullying. Combining machine learning models like BiGRU and CNN enhances the accuracy of detecting offensive and non-offensive tweets. However, there is a need to investigate how emerging deep learning models and related techniques can be combined to effectively detect and reduce cyberbullying on different platforms, including Twitter. Future studies should also investigate how models such as education and training can be improved to enhance the fight against pervasive cyberbullying. Future studies should review more recently published sources and identify the most accurate methods and techniques for reducing cyberbullying.

# **References**

Alotaibi, M., Alotaibi, B. and Razaque, A., (2021). A multi-channel deep learning framework for cyberbullying detection on social media. *Electronics*, *10*(21), p.2664.

Alsubait, T. and Alfageh, D., (2021). Comparison of machine learning techniques for cyberbullying detection on youtube Arabic comments. *International Journal of Computer Science & Network Security*, *21*(1), pp.1-5.

Cheng, l., Guo, R., Silva, YN, Hall, D. and Liu, H., (2021). Modelling temporal patterns of cyberbullying detection with hierarchical attention networks. *ACM/IMS Transactions on Data Science*, *2*(2), pp.1-23.

Cheng, l., Shu, K., Wu, S., Silva, Y.N., Hall, D.l. and Liu, H., (2020). Unsupervised cyberbullying detection via time-informed gaussian mixture model. *Arxiv preprint arxiv:2008.02642*.

Chua, YT, Parkin, S., Edwards, M., Oliveira, D., Schiffner, S., Tyson, g. and Hutchings, a., (2019). Identifying unintended harms of cybersecurity countermeasures. In *2019 APWG Symposium on Electronic Crime Research (ecrime)* (pp. 1-15). IEEE.

Cook, S., (2022). Cyberbullying facts and statistics for 2018 – 2022. *Comparitech.* [Accessed 1 November 2022], Available form: Https://www.comparitech.com/internet-providers/cyberbullying-statistics/

Davidson, T., Warmsley, D., Macy, m. and Weber, I., (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1, pp. 512-515).

Dixon, S., (2022). UK: harassment experienced online 2022. *Statista.* Https://www.statista.com/statistics/1319815/uk-online-abuse-experienced/

Faucher, C., Cassidy, W. and Jackson, M., (2020). Awareness, policy, privacy, and more: post-secondary students voice their solutions to cyberbullying. *European Journal of Investigation in Health, Psychology, and Education*, *10*(3), pp.795-815.

Gambäck, B. and Sikdar, U.K., (2017). Using convolutional neural networks to classify hate speech. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 85-90).

Gary W. Giumetti, Robin M. Kowalski (2022), Cyberbullying via social media and well-being, Current Opinion in Psychology, Volume 45, 101314, ISSN 2352-250X, https://doi.org/10.1016/j.copsyc.2022.101314. (https://www.sciencedirect.com/science/article/pii/S2352250X22000161)

Kaggle (2018). Tweets dataset for detection of cyber-trolls. *DATATUCKS.* Https://www.kaggle.com/datasets/dataturks/dataset-for-detection-of-cybertrolls

Kaluarachchi, C., Warren, M. and Jiang, F., (2020). Responsible use of technology to combat Cyberbullying among adolescents. *Australasian Journal of Information Systems*, *24*.

Khader, M., Karam, M. And Fares, H., (2021). Cybersecurity awareness framework for academia. *Information*, *12*(10), p.417.

Parkin, S. And Chua, YT, (2020). Refining the blunt instruments of cybersecurity: A framework to coordinate prevention and preservation of behaviours. In *International Workshop on Socio-Technical Aspects in Security and Trust* (pp. 23-42). Springer, Cham.

Perera, A. and Fernando, P., (2021). Accurate cyberbullying detection and prevention on social media. *Procedia Computer Science*, *181*, pp.605-611.

Quayyum, F., Cruzes, D.S. and Jaccheri, L., (2021). Cybersecurity awareness for children: A systematic literature review. *International Journal of Child-Computer Interaction*, *30*, p.100343.

Reddy, RC, Bhattacharjee, B., Mishra, D. and Mandal, a., (2022). A systematic literature review towards a conceptual framework for enablers and barriers of an enterprise data science strategy. *Information Systems and e-Business Management*, pp.1-33.

Singh, V.K. and Hofenbitzer, C., (2019). Fairness across network positions in cyberbullying detection algorithms. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 557-559). IEEE.

Talpur, BA and O'sullivan, D., (2020). Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in Twitter. In *Informatics* (Vol. 7, No. 4, p. 52). MDPI.

Waseem, Z. and Hovy, D., (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop* (pp. 88-93).