JOURNAL OF GOVERNANCE AND INTEGRITY
ISSN: 2600-7479 e-ISSN: 2600-786X
VOLUME 8, ISSUE 1, 2025, 874 - 880
DOI: https://doi.org/10.15282/jgi.8.1.2025.11508

RESEARCH ARTICLE

Social network approach for cyberbullying detection using machine learning

Anishah Muhammad Syafiq*, Mohd Faizal Ab Razak, Ahmad Firdaus Zainal Abidin, Salwana Mohamad@Asmara, Nur Khairunnisa Kamaruddin

Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan Campus, Pekan 26600, Pahang, Malaysia

ABSTRACT - Cyberbullying is a serious issue that affects both adults and teenagers on the internet. Using social media and the internet is frequently associated with sending, receiving, and publishing derogatory, false, or unpleasant content about other people. This shows that cyberbullying has had a substantial negative impact on mental health, especially among the younger population. If action is not taken to stop cyberbullying, self-esteem and problems with mental health will impact a whole generation of young adults. Considering this, it is necessary to use machine learning (ML) approaches combined with natural language processing (NLP) and techniques like TF-IDF (Term Frequency-Inverse Document Frequency) to detect cyberbullying effectively. This study utilised a dataset that includes data on hate speech tweets. The research began by analysing the nature of cyberbullying and the challenges in its detection, underscoring the significance of automated methods. The study used NLP and TF-IDF to pre-process and analyse the dataset, identifying patterns and characteristics typical of cyberbullying behaviours. Subsequently, various ML techniques were utilised to accurately train models that can detect instances of cyberbullying in social media content. Specifically, the study has three primary goals: to create and implement an effective method for detecting online abusive and bullying messages by integrating NLP with ML; to evaluate the accuracy of the proposed detection algorithms for cyberbullying, specifically using Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF); and to compare the performance of these algorithms to identify the one with the highest accuracy in detecting text-based bullying. To sum up, this study highlights the potential of leveraging ML, NLP, and TF-IDF to address the escalating issue of cyberbullying on social media. The study advances the development of sophisticated detection algorithms by utilising a comprehensive dataset to emphasise the need for a multimodal approach that combines technological solutions with awareness-raising and education to foster a safer and more inclusive online community.

ARTICLE HISTORY
Received : 04-11-2024
Revised : 12-11-2024
Accepted : 20-01-2025
Published : 30-07-2025

KEYWORDS
Fake news
Support vector machine
Naïve Bayes
Decision tree

1. INTRODUCTION

Online platforms have completely changed how we connect, communicate, and interact with one another, especially in a world where social media and digital technology prevail. However, there is a darker side to these virtual groups than meets the eye, as people indulge in dangerous behaviour behind their computers and in anonymity. According to a study (Pandey and Sharma, 2022), one of the most important problems to emerge in the internet age is cyberbullying. Cyberbullying is the deliberate and ongoing use of internet platforms to harass, threaten, or intimidate others, causing them to suffer psychologically or emotionally. Its effect can be crushing, mainly when it focuses on the youthful and powerless. Subsequent research by Peled (2019) indicates that cyberbullying leads to psychiatric disorders, including behavioural problems, along with physical and emotional harm to vulnerable victims. In short, cyberbullying is the intentional and ongoing harassment of someone online that significantly undermines their psychological and emotional well-being, especially among vulnerable youth.

According to Figure 1, The frequency of cyberbullying is alarmingly high, as 87% of middle and high school students reported having witnessed cyberbullying occurrences, and 36.5% reported having personally experienced it, according to a recent survey. Victims suffer grave repercussions, such as diminished academic performance, depression, and thoughts of suicide. Therefore, cyberbullying detection using a Machine Learning (ML) approach is suggested. The prevalence of cyberbullying on social media has prompted an urgent need for innovative solutions to detect and prevent this harmful behaviour. According to research (Jadhav et al., 2023), a model to identify cyberbullying can be created by using ML to identify language patterns used by bullies. In this regard, ML, a branch of artificial intelligence, conveys a promising approach to combat cyberbullying. Using ML approaches, reliable, effective, and scalable detection systems can thus be developed by automatically identifying patterns, language, and behaviours linked with cyberbullying. A reliable and diverse dataset is essential to harness the potential of ML for cyberbullying detection. For this study, we utilised the public Twitter dataset, sourced from the renowned platform and the academic publication "SOSNET: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection" (2020). The rise of cyberbullying calls for innovative solutions. ML, a subset of AI, offers a promising approach by identifying cyberbullying-related patterns and behaviours.

*CORRESPONDING AUTHOR | Anishah Syafiq | ✉ anishahsha99@gmail.com

According to Jadhav et al., models can be created to detect cyberbullying through language patterns. This study employs a publicly available Twitter dataset to create a machine-learning model for identifying cyberbullying, underscoring the feasibility of automated, scalable detection methods. This study aims to evaluate the accuracy of various algorithms in detecting abusive and bullying messages online. The study has three principal objectives: to develop and execute an efficient approach for identifying online abusive and bullying messages through the integration of natural language processing (NLP) and ML; to assess the precision of the proposed detection algorithms for cyberbullying, specifically employing SVM, Naive Bayes (NB), and Random Forest; and to compare the efficacy of these algorithms to ascertain the one with the most remarkable accuracy in detecting text-based bullying.

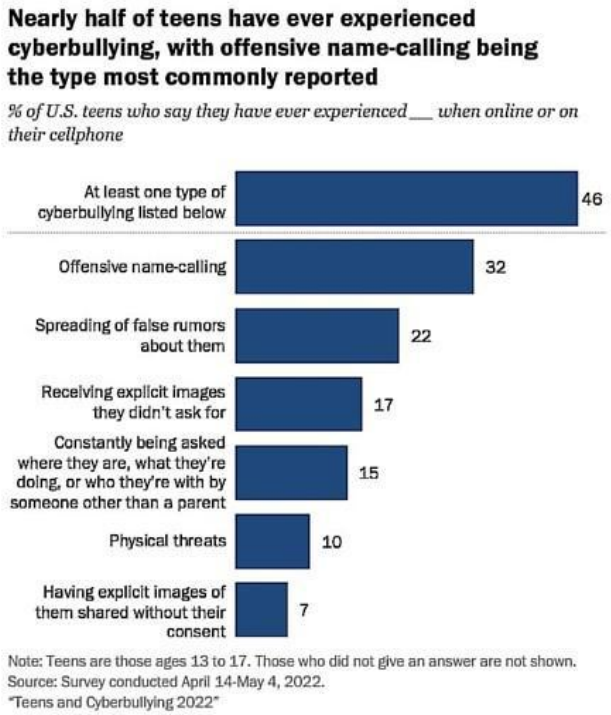

Figure 1. Example of cyberbullying

In conclusion, combating cyberbullying using ML, NLP, and TF-IDF techniques highlights the significance of governance in creating and implementing technology solutions. Robust governance frameworks guarantee the responsible and transparent utilisation of technologies to mitigate cyberbullying. The result entails the formulation of explicit data privacy and security protocols with the ethical application of ML models. Governance methods must incorporate oversight tools to oversee the deployment of these technologies and ensure adherence to applicable legislation and standards.

2. LITERATURE REVIEW

Several approaches propose systems that can accurately and automatically identify cyberbullying on social media. Table 1 illustrates the research conducted by Van Hee et al. (2015), which examines the challenges posed by social media in assessing online interactions, explicitly concentrating on cyberbullying as a case of cyber victimisation. Development and analysis of a corpus of Dutch social media posts, classification of fine-grained text relevant to cyberbullying, including threats and insults, and identification of specific actors (harasser, victim, bystander) are all part of the study. Proof-of-concept studies were carried out to identify cyberbullying occurrences and types automatically. Furthermore, it argued the necessity for automated detection tools for extensive social media surveillance and prompt intervention in dangerous circumstances. Perera and Fernando's (2021) research presents a methodology for employing supervised ML techniques, namely SVM and Logistic Regression, to detect and prevent cyberbullying on social media platforms. The algorithm emphasises topics such as racism, sexual content, and physical abuse, employing TF-IDF, N-grams, sentiment analysis, and profanity detection to identify cyberbullying accurately. With 75.17% accuracy, the system worked with a Twitter dataset to highlight context and intentional harm while addressing the changing nature of language. The simple interface facilitates easy input and categorisation, while updated versions prioritise enhancing accuracy and including additional user and network functionalities. Islam et al. (2020) address the detrimental consequences of social media usage, encompassing online abuse, harassment, cyberbullying, cybercrime, and online trolling. The research uses ML and NLP to create an efficient method for identifying abusive and bullying remarks sent online. The study employs Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) to evaluate the performance of four ML algorithms: Random Forest, Decision Tree, NB, and Support Vector Machines (SVM). Research employing Twitter and Facebook datasets revealed that SVM with TF-IDF features exhibited the highest detection accuracy for messages associated with cyberbullying.

Table 1. Comparison cyberbullying detection

Researcher	Algorithm	Dataset	Accuracy
(Van Hee et al., n.d.)	SVM, NLP	Collecting data from the social networking site Ask.fm (http://ask.fm)	78.5%
(Perera & Fernando, 2021)	TF-IDF and SVM	Twitter dataset	75.17% SVM got the higher accuracy
(Islam et al., 2020)	Bag-of-Word and TF-IDF. Random Forest (RF), Support Vector Machine, and Naive Bayes	Dataset Facebook and Twitter from Kaggle	78.5% SVM
(Al-Garadi et al., 2019)	SVM, NB, KNN, LR	Data extracted from SM websites.	77%

Finally, the present work uses big data analytics to forecast instances of cyberbullying on social media. It emphasises how extensive datasets and advanced machine-learning techniques can enhance predictive accuracy. The study seeks to establish a robust framework for early diagnosis and prevention by examining several algorithms and their effectiveness in identifying patterns of cyberbullying conduct. The Al-Garadi et al. (2019) study addresses the importance of timely interventions in mitigating the detrimental effects of cyberbullying on victims.

3. METHODOLOGY

Developing a cyberbullying detection system necessitates a sequence of systematic steps, each essential to the system's functionality. The proposed algorithmic framework provides a structured approach that emphasises the technical aspects of ML alongside the ethical considerations necessary for responsible application.

3.1 Research Framework

The development of an ML model began with the collection of a broad, labelled dataset that includes content related to cyberbullying. Based on Figure 2, this research used a publicly available Twitter dataset, as described at the 2020 IEEE International Conference on Big Data. Data pre-processing was the next critical step, involving cleaning, transforming, and organising raw data for analysis or model training. Pre-processing in NLP involves a series of steps to clean and prepare text data for analysis. These steps typically include tokenisation, which splits text into individual words or tokens, and lowercasing, which converts all characters to lowercase for uniformity. Punctuation was eliminated to prevent the classification of punctuation marks as distinct tokens, and frequently occurring words that offer minimal semantic value, such as "and," "the," and "is," were excluded as stop words.

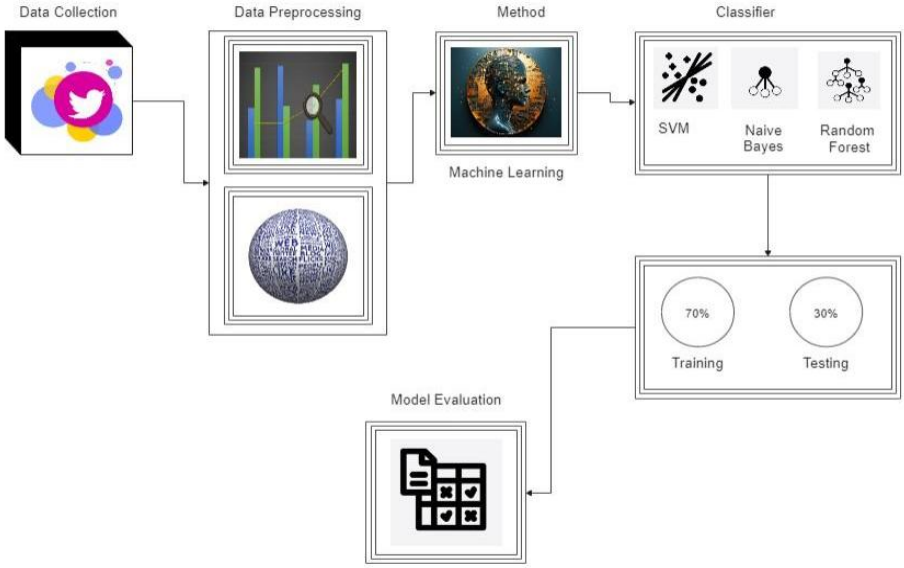

Figure 2. Research framework

Furthermore, stemming and lemmatisation served to condense words to their root forms. Stemming randomly removed word endings (e.g., "running" to "run"), while lemmatisation systematically transformed words to their dictionary form, considering context (e.g., "better" to "good"). Other pre-processing steps included removing numbers and special characters, performing spell correction, and normalising text to a standard format by removing accents or converting slang to its full forms. Together, these techniques helped reduce the dimensionality of text data and enhance the performance

of NLP models. The dataset was split for classification, with 70% used for training and 30% for testing. Choosing the appropriate ML algorithm, such as NB, SVM, or Random Forest classifiers, is essential for capturing the dynamics of cyberbullying. After classification, the model evaluation phase compared the accuracy of these classifiers to determine the best one for identifying cyberbullying content.

3.2 Project Requirement

The study aims to determine the best accuracy among three classifiers: NB, SVM, and Random Forest. Constraints encompass the necessity for comprehensive data filtration, the scrutiny of a considerable amount of data, the prolonged processing duration, and elevated RAM requirements for SVM. In the modern age of digital communication, social media platforms have become essential, but cyberbullying poses a severe threat to mental health. Traditional manual moderation is insufficient due to the volume of data, necessitating an automated, scalable solution. This study compares the three ML algorithms to identify cyberbullying on social media, using the public Twitter dataset for training and testing. The goal is to evaluate and rank these algorithms to determine the most accurate model for identifying cyberbullying content.

3.3 Proposed Design

The proposed design, illustrated in the flowchart, outlines integrating NLP with ML to create an effective method for detecting abusive and bullying messages online.

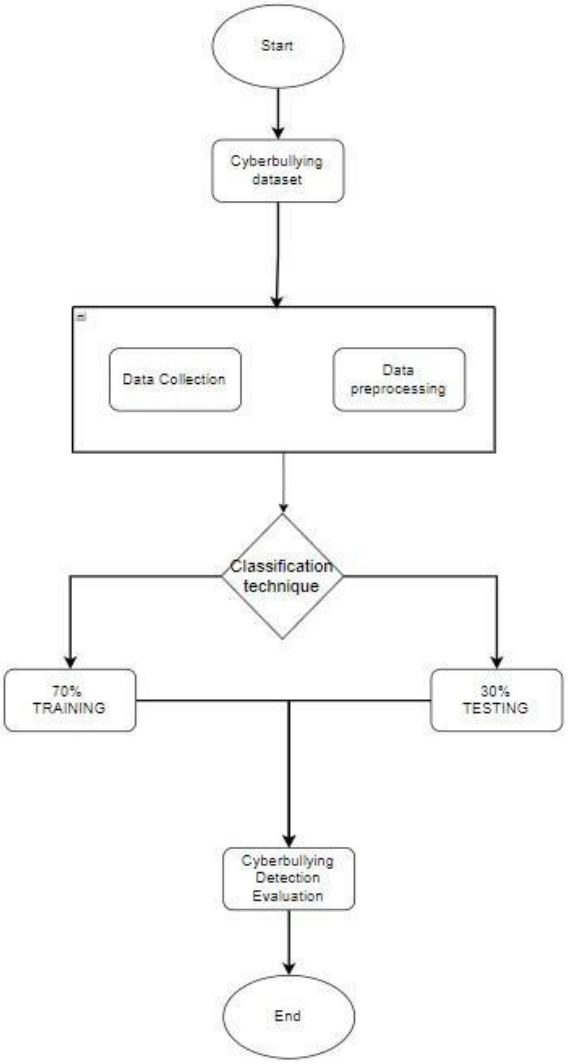

Figure 3. Flow chart

3.4 Data Design

The research utilised a public Twitter dataset from the 2020 IEEE International Conference on Big Data, consisting of around 47,000 tweets about cyberbullying. The dataset was categorised by age, gender, ethnicity, religion, and the presence of cyberbullying in the tweets, comprising 8,000 tweets per category. Upon collection, the tweets were classified and shown in a bar chart to illustrate the total frequency of each category of cyberbullying. This publicly available dataset may contain mistakes due to extensive data consumption. As such, confirming the null was crucial for ascertaining that the dataset is complete. The "isnull()" method was employed to identify missing values within the data frame, with each cell indicating either True (if the corresponding value is null) or False (if it is not). The outcome of testing the null data

is presented below. The boolean values of the data frame's columns—true represented as one and false as zero—were aggregated using the sum () function. The tweet_text and cyberbullying_type columns of the data frame contain complete values, as indicated by the output below.

Upon executing data.duplicated(), a boolean Series was produced, wherein each element was actual if the corresponding row was a duplicate of a preceding row and false otherwise. The.sum() method in this Series assigned a value of 1 to true and 0 to false when aggregating boolean variables. The total obtained was equivalent to the count of true values, as this count matched the number of duplicate rows. A word cloud visually represents term frequency, emphasising the most prevalent words linked to each category of cyberbullying. The results of the pre-processing method are presented below. The clean data that has undergone pre-processing were initially categorised by types of cyberbullying.

3.5 Hardware and Software

Before proceeding with the research, it is essential to establish a list of prerequisites outlining the hardware and software requirements for experimenting. The upcoming phase of this study involves rigorous testing and evaluation across both hardware and software platforms.

Table 2. Hardware and description

Hardware	Description
Processor: Intel® Core™ i7-4600U CPU @ 2.10GHz 2.70 GHz RAM: 16.0 GB - System type: 64-bit, Operating System, x64-based processor	Used to finish the resource seeking, implementation, testing, and documentation for the entire research project.

Table 3. Software and description

Software	Description
Anaconda: Jupiter Notebook	A data research programming language for Python that aims to streamline package management and deployment
Windows 10	A system for conducting this research
Microsoft 365 Microsoft words Microsoft Excel	To document the research To create planning and analysis for this research
OPERA	To gather and download the information.
Draw.io	To create most kinds of diagrams used in the documentation
Mendeley	To create references for information and insert citations into the thesis and report.

4. RESULTS AND DISCUSSION

Three distinct ML models—NB, Random Forest, and SVM—were employed to evaluate the classification problem. The NB classifier achieved an accuracy of 72.548%, with varying precision, recall, and F1 scores across different classes. While the model performed reasonably well, there were considerable disparities in recall and precision across numerous classes, suggesting potential areas for improvement. In comparison, the Random Forest and SVM classifiers exhibited superior performance relative to the others, achieving accuracies of 80.39% and 80.55%, respectively. These models exhibited enhanced memory, accuracy, and F1 scores across most classes, indicating a superior equilibrium between true positive and false positive rates. In certain classifications, the Random Forest model demonstrated significantly high precision and recall, signifying its effective identification of the underlying patterns within the dataset. The choice between Random Forest and SVM may be affected by various factors, including the application's specific requirements, interpretability, and computational efficiency. The results indicate the efficacy of group methods such as Random Forest and sophisticated algorithms like SVM in addressing classification issues, underscoring their potential application in real-world scenarios.

Table 4. Result accuracy

Model	Accuracy obtained
Support Vector Machine (SVM)	80.55%
Random Forest	80.39%
Naive Bayes	72.54%

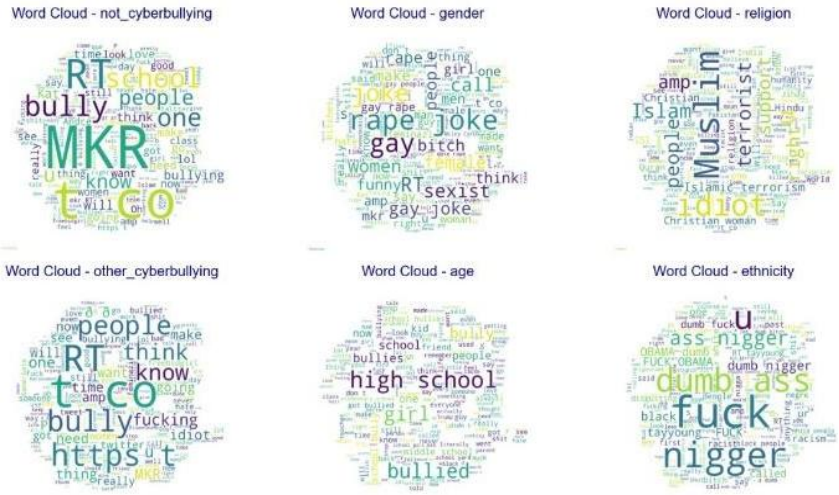

Figure 4. World cloud result

4.1 Model Evaluation

Model evaluation is critical in assessing the effectiveness of SVM, NB, and Random Forest models for classifying cyberbullying types from Twitter data. Accuracy, which measures the overall correctness of predictions, indicates the percentage of correctly classified instances out of the total predictions made by the model. Precision evaluates the proportion of correctly predicted positive instances (true positives) relative to all instances predicted as positive, highlighting the model's ability to avoid false positives.

$$Precision = \frac{true\ positive}{true\ positives + false\ positives}$$

On the other hand, recall assesses the proportion of correctly predicted positive instances out of all positive instances, emphasising the model's capability to identify all positive instances.

$$Recall = \frac{true\ positive}{true\ positives + false\ negatives}$$

The F1-score, which combines precision and recall into a single metric, provides a balanced assessment of a model's performance, instrumental when there is an uneven class distribution or when both precision and recall are crucial for the application.

$$F1 = 2.\ \frac{precision\ .recall}{precision + recall}$$

By employing these metrics comprehensively, one can effectively gauge and compare the performance of different models in accurately categorising cyberbullying instances, thereby enabling informed decisions in mitigating online harassment.

5. CONCLUSIONS

This research employs a ML technique to detect cyberbullying. Several ML techniques, including SVM, NB, and random forest, were proposed to develop a model and determine which yielded the highest accuracy. Evaluations were conducted to determine the optimal accuracy for each method. The results indicated that the SVM method achieved the highest accuracy, precision, recall, and f1 score, with an 80.55% outcome for each evaluation. Consequently, the research deduces that SVM is the most effective method for identifying cyberbullying.

ACKNOWLEDGEMENT

The authors would like to thank the Ministry of Higher Education for providing financial support under Universiti Malaysia Pahang Al-Sultan Abdullah for laboratory facilities and additional financial support under Internal Research grant RDU210321.

AUTHORS CONTRIBUTIONS

Anishah Muhammad Syafiq (Conceptualization; Methodology; Data curation; Formal analysis; Software; Visualization; Writing – original draft; Writing – review & editing)
Mohd Faizal Ab Razak (Supervision; Project administration; Guidance)
Ahmad Firdaus Zainal Abidin (Supervision; Guidance)
Salawana Mohamad @ Asmara (Supervision; Guidance)
Nur Khairunnisa Kamaruddin (Support)

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

ETHICAL STATEMENT

Not applicable.

CONFLICT OF INTEREST

The author(s) declare there are no conflicts of interest, financial or non-financial, that could have influenced the content of this manuscript.

REFERENCES

2020 IEEE International Conference on Big Data (Big Data). (2020). IEEE.

Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., Mujtaba, G., Chiroma, H., Khattak, H. A., & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. *IEEE Access, 7*, 70701–70718.

Ali, A., & Syed, A. M. (2020). Cyberbullying detection using machine learning. *Pakistan Journal of Engineering and Technology*, *3*(2), 45-50.

Islam, M. M., Uddin, M. A., Islam, L., Akter, A., Sharmin, S., & Acharjee, U. K. (2020, December). Cyberbullying detection on social networks using machine learning approaches. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (pp. 1-6). IEEE.

Modem, S., Girish, H., Prathap, J. A., Ramakrishnappa, M., & Student, U. G. (2023). Faults identification in digital counter using Naive-Bayes' algorithm. *Tuijin Jishu/Journal of Propulsion Technology*. *44*(4), 1145-1151.

Muneer, A., & Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, *12*(11), 187.

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

Simões, R. D. M., Huber, P., Meier, N., Smailov, N., Füchslin, R. M., & Stockinger, K. (2023). Experimental evaluation of quantum machine learning algorithms. *IEEE Access*, *11*, 6197-6208.

Pandian, A. P., Palanisamy, R., & Ntalianis, K. (Eds.). (2021). *Proceedings of International Conference on Intelligent Computing, Information and Control Systems: ICICCS 2020*. Springer.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., ... & Hoste, V. (2015). Automatic detection and prevention of cyberbullying. In *International Conference on Human and Social Analytics (HUSO 2015)* (pp. 13-18). IARIA.

Xu, B., Guo, X., Ye, Y., & Cheng, J. (2012). An improved random forest classifier for text categorization. *Journal of Computational Science*, *7*(12), 2913-2920.