Machine Learning-Based Detection of Cyberbullying in Social Media Texts: A New Zealand Contextual Study

Oras Baker
Faculty of Computing, Ravensbourne University London, SE10 0EW, United Kingdom

Kasthuri Subaramaniam*
Department of Decision Science, Faculty of Business and Economics, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

Sellappan Palaniapan
Help University, Bukit Damansara 50490 Kuala Lumpur, Malaysia

Dobrila Lopez
School of Computing, Eastern Institute of Technology, Auckland, New Zealand

Abdurrahman Bin Jalil
Faculty of Information Science, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia

Email: O.alhassani@rave.ac.uk, s_kasthuri@um.edu.my, sellappan.p@help.edu.my, DLopez@eit.ac.nz, abdurrahman.jalil@uitm.edu.my
*Corresponding Author

Abstract

This study investigates the use of machine learning approaches to identify cyberbullying within social media text, focusing on New Zealand. The research develops and evaluates supervised machine learning models trained on a labelled dataset of X (Twitter) posts obtained from Kaggle. Text preprocessing involved data cleaning, tokenisation, lemmatisation using WordNetLemmatizer, and vectorisation through Term Frequency–Inverse Document Frequency (TF–IDF). Several algorithms, including Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, and Random Forest, were implemented and compared using performance metrics. A Streamlit-based web application was developed to operationalise the most effective model, allowing users to detect cyberbullying in real time.

Keywords: Machine learning, Cyberbullying, Social media, Support vector machine

1. Introduction

In this present era of advanced technology, social media platforms have become an important part of our lives, connecting us with friends, family, and the wider world. According to Zainol et al. [1], "social media and "always on always-connected devices" have completely changed the phase of social interactions". They have transformed the way we communicate, share information, and stay updated with current events. However, the use of applications also has consequences, which can lead to some aggravated aggressive and bullying behaviours among users. The anonymity and scope of social networks can put the ill in people which can result in destructive interactions, with significant consequences. Cyberbullying, which is defined as "any behaviour conducted through electronic means by people or groups of individuals that regularly sends hostile or violent messages designed to inflict injury or discomfort on others" [2], has been rising to alarming levels of concern among individuals in this digital environment.

According to Rosa et al. The increase in access to information and communication technologies and increased use of computers and smartphones for children and teenagers worldwide has made cyberbullying an international problem [3]. Considering that online harassment is so prevalent, we need measures that recognize and penalize harassment online, especially on sites like Twitter. Cyberbullying also poses a significant problem regarding the impact on young people such as the predominant group in young users of social media in New Zealand [4]. Online interaction is anonymous, such as social media, these nefarious behaviours are less easily visible and thus, it is harder to make offenders and victims take accountability for wrongdoing [5].

Text classification works from studies in New Zealand have revealed that fine preprocessing (tokenisation, lemmatisation, TF–IDF), balanced datasets and evaluation

© The 2026 International Conference on Artificial Life and Robotics (ICAROB2026)

of comparative algorithms (SVM, NB, LR, RF) lead to a strong recognition of cyberbullying, robust detection [6], [7]. Twitter sentiment analysis and pandemic research show both the difficulties of adaptation at the domain level and annotation needs. Usable real-time interfaces are supported by examples from web deployments and by HCI literature, and recommender/optimisation models can assist in sampling and feature selection with respect to detections for recall and precision improvements [8], [9], [10], [11].

Nevertheless, there may be several challenges in employing strong machine learning models for its application to cyberbullying detection. The complexity of human speech, including slang and cultural references, and the use of diverse languages, such as in New Zealand, pose a challenge on the part of models to accurately predict the cases of cyberbullying. Moreover, benign content that appears benign with a false positive label, or harmless (false negative) as harmful could yet also find false negative detection; and there is the danger of being overlooked by the data itself. Therefore, while machine learning may improve identification of cyberbullying on platforms like Twitter, there is the need to develop, train and evaluate models carefully to guarantee their accuracy and neutrality.

2. Methodology

A structured data collection and preprocessing process will be conducted. This research will be based on a labelled dataset of social media posts collected from Kaggle. Raw text will be preprocessed using cleaning, tokenization, lemmatization with NLTK's WordNetLemmatizer, and vectorization using Term Frequency-Inverse Document Frequency (TF-IDF). This preprocessing pipeline will transform the textual data into a numerical format suitable for machine learning model ingestion.

Following data preprocessing, the selected machine learning models will be trained on the prepared dataset. The best model for cyberbullying detection will be determined by carefully assessing the model's performance using measures including accuracy, precision, recall, and F1-score. Particular emphasis will be placed on addressing challenges associated with imbalanced datasets and optimising model performance.

The Kaggle dataset is quite useful for this research because most of the data sets on platforms use data obtained from real people's posts. In contrast to controlled or simulated data, the data gathered from these platforms reflect the nature of any other social media communication [12]. They are informal, truncated, and often rather noisy. The content of the social media posts involves multiple slangs and low standard English, abbreviations, and emojis, and improper grammar that makes challenges as to how to feed it most properly to the machine learning models for best predictions. Also, this dataset can cover all manners of cyberbullied, ranging from verbal to explicit abuse, which will help to train the detection system rather effectively.

One of the biggest advantages of this kind of dataset is that it contains a variety of linguistic forms while maintaining explicit labels that provide good conditions for feature extraction. For instance, some entries containing such words and phrases as loser, idiot, hate, etc. can be considered to contain evidence of cyberbullying. Minimally supervised learning enables the machine learning models to use the patterns between words and the context in which they appear to accurately determine further patterns of detrimental conduct [13].

3. Analysis and Results

This section will demonstrate how the locally trained SVM model works within the Streamlit interface and analyse the performance as well as the response of that model to different text inputs. The objective of testing is to review how the application performs under real-life conditions and check if it can offer identical consequences constantly.

The testing process involved evaluating the app by inputting different types of text samples, including:

- Non-cyberbullying text: Friendly or neutral messages to test the model's ability to correctly identify non-offensive content.
- Cyberbullying text: Offensive or harmful messages to test the model's capacity to detect cyberbullying.

The text was tested through the Streamlit interface by entering the samples into the text input box and pressing the Classify button. The model's prediction appeared on the screen immediately after processing the input, providing real-time feedback.

(Figure 1) shows case 1 of non-cyberbullying text.

Input: "You are doing an amazing job, keep up the great work!"
Prediction: "Not Cyberbullying"



Figure 1 Sample test case on Streamlit showing the result for cyberbullying case 1.

(Figure 2) shows case 2 of cyberbullying text.

Input: "You're such a loser, nobody likes you."
Prediction: "Cyberbullying"

© The 2026 International Conference on Artificial Life and Robotics (ICAROB2026)

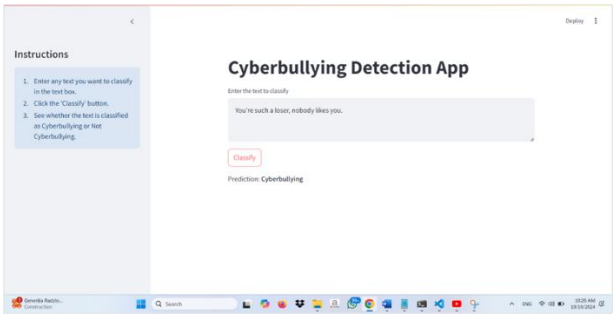


Figure 2 Sample test case on Streamlit showing the result for cyberbullying case 2.

It is vital to optimise the cyberbullying detection app so that the text processing runs smoothly and results are predictable. The SVM (Support Vector Machine) model, which is also central to the application, is one of its core areas needing optimisation. Even though the SVM model performed well at the start of testing, there are many ways to improve its performance:

- Hyperparameter Tuning - As the SVM model currently uses default hyperparameters not always optimised for the dataset. Hyperparameter tuning, such as grid search or random search, would help in finding the optimal parameters (e.g., C, gamma) that maximise the model's accuracy, precision, and recall. Fine parameter tuning in turn enables the model to perform better in edge cases, i.e., detect covert cyberbullying.
- Model Selection - SVM had robust results, but trying to test ensemble models like Gradient Boosting and XGBoost may give better results. These models usually do well in complex patterns and imbalanced datasets which will ultimately limit false positives and negatives.
- Training in Larger Datasets - More data enhances the performance of machine learning models. If we want to improve the accuracy and generalisability of the SVM model, we could retrain the SVM model on a larger dataset (e.g., collect more labelled cyberbullying examples from Twitter) to understand more of the text patterns involved and improve its predictive effect.

The quickness of the application is important for good user experience. While Streamlit is built for rapid deployment, additional optimisations are proposed to minimise the time in the process from user input to model predictions, such as:

- Minimising Preprocessing Time - The present preprocessing steps (cleaning, tokenization, lemmatization) are efficient but with improved methods it would be time-effective for the user input to be preprocessed in less time. For example, cached preprocessing pipelines would speed up the computation of many words and phrases.
- Efficient Model and Data Loading - The cache process for the model and vectorizer are stored in memory and it means loading these objects is minimised, and the application won't load them every time it starts. Streamlit includes storage for caching (@st.cache) that loads commonly used data and models once and thus the app's response time is shorter.

4. Conclusion

Finally, we successfully proved the applicability of machine learning-based models over tweets to detect cyberbullying, in particular Support Vector Machines (SVM), under Twitter's situation. On the basis of Natural Language Processing (NLP) techniques such as tokenization, lemmatization, and TF-IDF vectorization, the models accurately classified tweets as cyberbullying or non-cyberbullying and were mainly led by SVM. The study also addressed some of the major aspects like dataset imbalance and noisy textual data, while solving problems like resampling techniques to greatly improve performance of the models.

This research has important implications as this study is related with the development of safer online environments. By integrating such machine learning models into social media platforms, harmful content will be caught and remedied much more quickly, to reduce the impact of cyberbullying on vulnerable users. But implementation of these models in real-world use needs to be further optimized, mainly in the areas of real-time detection and the analysis of multilingual data.

The significance of automated applications for content moderation and monitoring will only increase as social media plays a pivotal role in ensuring global connections.

References

1. Z. Zainol, S. Wani, P. N. Nohuddin, W. M. Noormanshah, and S. Marzukhi, "Association analysis of cyberbullying on social media using Apriori algorithm," Int. J. Eng. Technol., vol. 7, no. 4.29, pp. 72–75, 2018.
2. K. Amran and M. S. Hassan, "A review on cyberbullying prevention in social media among adolescents," Int. J. Acad. Res. Bus. Soc. Sci., vol. 13, no. 3, 2023.
3. H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho, "A 'deeper' look at detecting cyberbullying in social networks," in Proc. 2018 Int. Joint Conf. Neural Netw. (IJCNN), 2018, pp. 1–8.
4. M.-J. Wang, K. Yogeeswaran, N. P. Andrews, D. R. Hawi, and C. G. Sibley, "How common is cyberbullying among adults? Exploring gender, ethnic, and age differences in the prevalence of cyberbullying," Cyberpsychol., Behav., Soc. Netw., vol. 22, no. 11, pp. 736–741, 2019.
5. K. Subaramaniam, R. Kolandaisamy, A. B. Jalil, and I. Kolandaisamy, "Cyberbullying challenges on society: A review," J. Positive School Psychol., vol. 6, no. 2, 2022.
6. O. Baker, J. Liu, M. Gosai, and S. Sitoula, "Twitter sentiment analysis using machine learning algorithms for COVID-19 outbreak in New Zealand," in Proc. 2021 IEEE 11th Int. Conf. System Eng. Technol. (ICSET), 2021, pp. 286–291.
7. O. Baker, W. Li, and D. Liu, "Advanced web optimization: Leveraging neural collaborative filtering and prefetching

© The 2026 International Conference on Artificial Life and Robotics (ICAROB2026)

for enhanced user responsiveness," in Proc. IEEE 7th Int. Conf. Electrical, Electronics and System Eng. (ICEESE), 2024, pp. 1–6.
8. K. Subaramaniam and O. F. Baker, "Human personality types and software interface design: HCI from a different perception," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 1, no. 3, pp. 253–256, 2011.
9. O. Baker, Z. Ziran, M. Mecella, K. Subaramaniam, and S. Palaniappan, "Predictive modelling for pandemic forecasting: A COVID-19 study in New Zealand and partner countries," Int. J. Environ. Res. Public Health, vol. 22, no. 4, p. 562, 2025.
10. O. Baker, W. Li, and Q. Yuan, "Advanced human motion detection and precision movement measurement via mobile devices using YOLOv8, R-CNN, and augmented reality," in Proc. 2024 Int. Symp. Networks, Computers and Commun. (ISNCC), 2024, pp. 1–6.
11. K. Subaramaniam, J. L. Ern-Rong, and S. Palaniappan, "Interface design with personality types: An effective e-learning experience," EVERGREEN – Joint J. Novel Carbon Res. Sci. & Green Asia Strategy, vol. 8, no. 3, pp. 618–627, Sep. 2021.
12. Y. K. Cheah and O. Baker, "Location-based mobile augmented reality application for tourism," in Proc. 2020 IEEE Graphics and Multimedia (GAME), 2020, pp. 37–42.
13. X. Guo and S. Gauch, "Using sarcasm to improve cyberbullying detection," in Proc. Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING, 2024, pp. 52–59.

Authors Introduction

Dr. Oras Baker



He is an Associate Professor and Head of Masters in Cyber Security and Cyber Security Management at University of Ravensbourne London, UK. With 25 years of distinguished experience spanning academia, research, and industry, he specialises in Artificial Intelligence, Software Engineering, Cyber Security, Data Mining, and Machine Learning.

Ts. Dr. Kasthuri Subaramaniam



She received her Bachelor's and Master's degrees in Computer Science from Universiti Malaya. She earned her Ph.D. in Informatics from Malaysia University of Science & Technology. Her research interests include human-computer interaction, artificial intelligence and machine learning. She is a Senior Member of IEEE.

Dr. Sellappan Palaniappan



He is a Professor at Help University, Malaysia. He received the Ph.D. degree in Interdiscplinary Information Science from Universiti of Pittsburg. He has more than 40 years of teaching experience, authored several IT/CS books, published more than 100 journal/ conference papers and keynote speaker in conferences His research areas include machine learning, artificial intelligence, data science and cybersecurity.

Dr. Dobrila Lopez



She is a principal lecturer and postgraduate coordinator at the Eastern Institute of Technology, Auckland, New Zealand. She earned her PhD from Curtin University, Australia and her MEng from University of Auckland, Engineering School. Her research interests include technology in education, machine learning and cyberbullying.

Dr. Abdurrahman Bin Jalil



He earned his Bachelor of Science in Information Study from Universiti Teknologi MARA in 2001, completed his Master of Science in Computer Science in 2004 from Universiti Teknologi Malaysia and received his PhD at Sheffield Hallam University in the United Kingdom. His research interests include mobile application development, information systems, and artificial intelligence.

© The 2026 International Conference on Artificial Life and Robotics (ICAROB2026)