Investigating The Efficiency of ML Algorithms in Detecting Cyberbullying on Social Media

Vedant Marwadi - NXJ4679 [1]

[1]Master Of Computer And Information Science , Cybercrime and Cybersecurity, Auckland University of Technology

Abstract

The number of social media users is on the rise. It has been observed that social media platforms adjust based on user preferences and behaviours, making them more addictive. Whereas social media gives advantages, it additionally presents hazards such as cyberbullying. This paper investigates how functional ML algorithms are in identifying cyberbullying in multiple languages on social media as well as in detecting sarcasm, hate speech, aggressive comments and bully images, by reviewing 25 research studies on the same subjects. It has been found that RF was one of the algorithms used that achieved 99.99% accuracy in detecting cyberbullying. Besides, ML algorithms range from 75% to 98.82% accuracy in detecting cyberbullying in different languages including English, Hinglish, Hindi, Malay, Roaman Urdu, Arabic, Bengali, Bangla and Chinese. CNBD was 98.23% precise in locating bully images. SVM and Ensemble model was 79% correct in detecting sarcasm. While ANN was 92.9% accurate in finding hate speech, RoBERta achieved 0.87 F1 score in detecting aggression in comments. Overall, it was found that ML offers a valuable tool for combating cyberbullying. This research opens doors for developing more robust and responsible ML solutions to create a safer online environment

Keywords: Machine Learning, Cyberbullying, Social Media, Multilingual, Detection

1. Introduction

In this era of 21 century the usage of social media has peaked enormously. According to current GWI research, an average individual uses social media for two hours and twenty-three minutes per day. (Kemp, 2024). Ostic described social media as a platform that allows adolescents to communicate with each other through various electronic communication technologies, such as text messages, comments, and digitally altered images and videos (Ostic et al., 2021).

The population of the world increased over the years and so did the number of social media users. Throughout 2023, 266 million new individuals created their first-ever social media account. This staggering number additionally demonstrates that during the year prior, the world got average 8.4 new social media head per second. (Kemp, 2024).

While social media offers benefits like expanding social networks and creativity outlets for teens, it also poses risks such as cyberbullying (Allen, 2019). The anonymity and public nature of social media contribute to the rise of cyberbullying. It is true that bullying existed even before the world of internet however the proliferation of digital platforms and the surged in smartphones have led to an environment where some people can bully other individuals over the cyberspace. (Bergman, 2024).

With the increasing number of social media users, orthodox methods of detecting cyberbullying on social media are falling behind. On the other hand, ML algorithms are shown to be more effective. The intent of this research is to look into the usefulness of ML algorithms. It also delves into how efficient they are in different languages in detecting sarcasm, as well as in spotting bully images, hate speech, and aggressive comments on social media, by reviewing other papers in the same fields.

2. Literature Review

2.1 Usefulness of ML Algorithms in Detecting Cyberbullying on Social Media

Muneer and Fati discussed the importance of finding out and precluding cyberbullying on media platforms using ML algorithms and used the dataset of 37,373 tweets to evaluate seven common classifiers. Out of the seven classifiers, LR achieved the best accuracy of 90.57% and an F1 score of 0.9280 (Muneer & Fati, 2020).

Identically, Alqahtani and Ilyas utilized a multiclass dataset of cyberbullying tweets publicly available on Kaggle. The paper involved the development of an ensemble model that combined three multi-classification models (Decision Tree, Random Forest, and XGBoost) to detect cyberbullying in tweets. The ensemble models, employing both voting and stacking techniques, were evaluated and exhibited an accuracy rate of 90.71% (Alqahtani & Ilyas, 2024).

Furthermore, Tapsoba decided to investigate the role of ML algorithms in enhancing the identification of cyber threats, specifically focusing on phishing, cyberbullying, and online scams. The data for cyberbullying was collected from over 47,000 tweets. The research yielded encouraging outcomes in spotting cyberbullying using ML algorithms.

Random Forest appeared as the most effective model, achieving exceptional accuracy rates of 99.99% in cyberbullying detection (Tapsoba et al., 2024).

Lastly, Xingyi and Adnanr determined to develop a cyberbullying detection framework that incorporated the BERT pre-training model for word embedding, the BiSRU++ model with attention mechanisms for contextual feature extraction, and multi-task learning for joint training of sentiment analysis. Compared to traditional rule-based and statistical-based models, the proposed framework achieved an perfection of 86.1% in terms of the F1 score (Xingyi & Adnan, 2024).

2.2 Efficiency of ML Algorithms in Detecting Cyberbullying in Different Languages on Social Media

Raj aimed to use deep learning techniques to progress a model which can dredge up cyberbullying in tweets in multiple languages and produced CNN-BiLSTM model made of CNN and a BiLSTM network. It achieved accuracy of 94.94% in detecting cyberbullying for the dataset used, which contained Hindi, English, and Hinglish (a mixture of Hindi and English) languages (Raj et al., 2022). On the other hand, Singla focused on detecting cyberbullying specific for Hinglish language and collected Facebook comments for the dataset. Five machine learning algorithms were used in the study, out of which SVM achieved the highest accuracy of 87.53% (Singla et al., 2023).

Detecting cyberbullying in Malay was the focus of Ismail. To achieve this, 165,239 real-world comments connected to 27 public Instagram profiles of well-known Malaysian celebrities were gathered and stored in a dataset. The Support Vector Machine SVM method was employed, and it proved to be 75% accurate in categorizing comments related to cyberbullying (Ismail et al., 2024).

Dewani set out to find instances of cyberbullying aimed toward Roman Urdu. The Roman Urdu comments from social media platforms were collected to construct the dataset that RNN-LSTM, RNN-BiLSTM, and CNN were applied to. With reasonable accurateness of 85.5% and 85%, individually, RNN-LSTM and RNN-BiLSTM fared better than the CNN model (Dewani et al., 2021).

AlHarbi proposed to track down cyberbullying in Arabic social media. RR and LR are suggested as the best approaches for cyberbullying detection (AlHarbi et al., 2020). For the same purpose, Alduailaj & Belghith used Support Vector Machine (SVM). The paper also included the Farasa tool to enhance text analysis, which led to a better identification of cyberbullying. When combined with TF-IDF and BoW which are feature extraction approaches, the SVM classifier was able to recognize Arabic cyberbullying tweets with a high accuracy rate of 95.742% (Alduailaj & Belghith, 2023).

Akhter utilized four ML algorithms to identify cyberbullies on social media in Bengali language. This study used both binary and multilabel classifications; for the binary categories, Logistic Regression produced accuracy rates of 98.57%, while for the multilabel classifications, Multilayer Perceptron produced accuracy rates of 98.82% (Akhter et al., 2023).

Similarly, Chen proposed a mixed detection model based on XLNet and deep Bi-LSTM for exposing cyberbullying in Chinese. The F1-score of the alleged model reached 90.43% on Chinese offensive language dataset (Chen et al., 2024).

Closely, Mahmud focused on detecting vulgar language in Bangla and especially in Chittagonian dialect. The paper concluded that logistic regression, and simpleRNN were effective with over 90% accuracy (Mahmud et al., 2023).

2.3 Applicability of ML Algorithms in Detecting Sarcasm in Cyberbullying Text on Social Media

These days, it's common to witness cyberbullies sending sarcastic remarks. Ali & Syed mentioned that sarcasm is oftentimes used as an subtle and deceptive form of bullying and can have adverse effects on the victim. The aim of the paper was to detect sarcasm as an essential aspect of cyberbullying detection. For that, it used four different datasets and applied ML algorithms such as SVM, naïve Bayes, Random Forest, Logistic Regression. Furthermore, it also used ensemble approach which was a hybrid model according all the algorithms mentioned before. The outcomes delineated that SVM and Ensemble carried out better than the abiding classifiers with 79% average accuracy. (Ali & Syed, 2020).

2.4 Relevance of ML Algorithms in Detecting Bully Images on Social Media

Jadhav concentrated on determining cyberbullying in not only text but images as well. For that reason, it used the dataset from GitHub comprising 500 training and 50 testing images. It employed MobileNetV2, a CNN, for image based cyberbullying detection, which terminated the accurateness of 86% (Jadhav et al., 2023).

For the same purpose, Pericherla Ilavarasan introduced a deep learning method called CNBD, which stands for Combinational Network for Bullying Detection that combined the BEiT and MLP network and used IC and OCR to extract text features from images. In order to detect bully images, this paper collected 19,300 images from various social media platforms, and the model was reported to be 98.23% accurate (Pericherla & Ilavarasan, 2024).

Correspondingly, Almomani shared the same aim as the previous two research. Two experiments were conducted in the paper. The first experiment was to train the model over 5 epochs, and the other was to extend the training to 20 epochs. Five pre-trained models, VGG16, VGG19, ResNet50, InceptionV3, and InceptionResNetV2, were utilized for both the experiments in which InceptionV3 attained the upmost preciseness of 65% in the 5 epoch experiment, while ResNet50 performed best with a 63% accuracy in the 20 epoch experiment. Furthermore, the study also proposed a hybrid model that combined DL models as feature extractors with traditional ML classifier. The hybrid model achieved a solid accuracy of 82% (Almomani et al., 2024).

2.5 Functionality of ML Algorithms in Detecting Hate Speech on Social Media

Yuan concentrated on hate speech detection in social media platforms and presented the HateNet model that saddled deep neural network architecture and t-HateNet model which was an augmentation of the HateNet model that used transfer learning strategies. Of the two datasets utilized in the examination, t-HateNet model was 75.78% exact on the Davidson dataset and 77.48% on the Waseem dataset (Yuan et al., 2023).

Sultan involved shallow as well as DL techniques for a similar reason. Out of the nine algorithms employed in the paper, three were DL and six were shallow learning. The paper expressed that while DL methods automatically learn features from data, shallow ML methods often require manual feature engineering. Therefore, CNN achieved the highest accuracy of 90.2% in detecting hate speech in the Hate Speech and Offensive Language Dataset, while shallow machine learning algorithms exhibited accuracy rates ranging from 60.2% to 87.4% (Sultan et al., 2023).

In addition to that, James & Osubor developed a machine learning framework using an Artificial Neural Network (ANN) to filter anti-female jokes in order to detect harassment on social media. The framework effectively classified these jokes with a high performance accuracy rate of 92.9% on the dataset, which comprised 312 one-liner humorous jokes with attributes related to human centeredness and polarity orientation (James & Osubor, 2022).

2.6 Practicality of ML Algorithms in Detecting Aggressive Comments on Social Media

Paul shed light on the prevalence of aggressive content on social media platforms and reviewed various research studies in this area. The research examined papers that used a lexicon based approach as well as different ML and DL algorithms and implied that these

algorithms could indeed be successful in detecting aggressive content on social media platforms however the accuracy varies depending on the factors such as specific algorithm used, the kind of training data, and features extracted from text data. Furthermore, the research paper also suggested that SVM had been extensively used and found effective in identifying aggressive matter on different social media websites (Paul et al., 2020).

Additionally, Ejaz informed that existing datasets often focus only on aggressive texts, neglecting other crucial aspects of cyberbullying, and proposed to develop a more exhaustive dataset that included multiple aspects such as aggressive texts, repetition, peerness, and intent to harm. Additionally, it compared the interpretation of traditional ML algorithms and DL models and used specific thresholds for the mentioned elements to detect cyberbullying in text messages in which the RoBERTa Base model trained on the dataset achieved the best performance, with an F1-score of 0.87 and among traditional ML models, the SVM accomplished the best execution with the utmost F1-score of 0.83 (Ejaz et al., 2024).

3. Findings

Table 1

Author (Year)	**Language**	**Algorithm**	**Accuracy (%)**
Raj et al. (2022)	English, Hindi, Hinglish	CNN-BiLSTM	94.94
Singla et al. (2023)	Hinglish	SVM	87.53
Ismail et al. (2024)	Malay	SVM	75
Dewani et al. (2021)	Roman Urdu	RNN-LSTM/BiLSTM	85.5
AlHarbi et al. (2020)	Arabic	Logistic Regression/Ridge Regression	-
Alduailaj & Belghith (2023)	Arabic	SVM	95.74
2*Akhter et al. (2023)	Bengali (Binary)	Logistic Regression	98.57
	Bengali (Multilabel)	Multilayer Perceptron	98.82
Chen et al. (2024)	Chinese	XLNet-BiLSTM	90.43
Mahmud et al. (2023)	Bangla (Chittagonian)	Logistic Regression/SimpleRNN	>90

Table 1 depicts efficiency of different ML algorithms in detecting cyberbullying in various languages on social media. The languages are English, Hindi, Hinglish, Roamn Urdu, Arab, Bengali, Bangla and Chinese. The accuracy of these algorithms ranges from 75% to 98.82%. This table convey proof that ML calculations can be effective in finding cyberbullying over various languages. At the same time, the most productive algorithm depends on the particular language and the constitution of the training data. Future study straightly matching algorithms on standardized datasets and tasks would be advantageous for definative conclusions.

Table 2 illustrates that both ML as well as DL algorithms are useful in detecting cyberbullying having average accuracy of 93.76%. The chart also reveals that CNDB model obtained the greatest accuracy of 98.23% in spotting bully photos on social media. Furthermore, SVM, Naive Bayes, Random Forest, Logistic Regression and Ensemble model were used to detect the element of sarcasm in cyberbullying in which SVM and

Table 2

Author (Year)	Model	Accuracy (%)
Cyberbullying Text Detection		
Muneer & Fati (2020)	Logistic Regression (LR)	90.57
Alqahtani & Ilyas (2024)	Ensemble Model (Voting/Stacking)	90.71
Tapsoba et al. (2024)	Random Forest (RF)	99.99
Xingyi & Adnan (2024)	BERT + BiSRU++ with Multi-task Learning	86.1 (F1-score)
Bully Image Detection		
Jadhav et al. (2023)	MobileNetV2 (CNN)	86
Pericherla & Ilavarasan (2024)	CNBD (BEiT + MLP) with Image Captioning & OCR	98.23
2*Almomani et al. (2024)	Pre-trained models (VGG16, VGG19, ResNet50, InceptionV3, InceptionResNetV2)	InceptionV3 (65) & ResNet50 (63)
	Hybrid model (Deep learning + Machine learning)	82
Sarcasm Detection in Cyberbullying Text		
5*Ali & Syed (2020)	SVM	79
	Naive Bayes	76
	Random Forest	76.70
	Logistic Regression	78
	Ensemble	79
Hate Speech Detection		
2*Yuan et al. (2023)	HateNet	75.78
	t-HateNet	77.48
2*Sultan et al. (2023)	CNN (Deep Learning)	90.2
	Various Shallow Learning Algorithms	60.2 - 87.4
James & Osubor (2022)	ANN	92.9
Aggressive Comment Detection		
Paul et al. (2020)	SVM	-
2*Ejaz et al. (2024)	RoBERTa	87 (F1 score)
	SVM	83 (F1 score)

Ensemble approach performed the best with the accuracy score of 79%. While accuracy of algorithms used to detect hate speech on social media platforms range from 60.2% to 92.9%, algorithms used to detect aggressive text achieved average F1 score of 85%.

4. Conclusion And Discussion

Based on the findings it can be concluded that AI can detect cyberbullying on social media in multiple languages. Furthermore, ML algorithms even show good precision in detecting bully images. They are also proven to identify sarcasm, hate speech and aggression in comments as well. Hence, we can say that AI detection algorithms can be the antidote to curbing cyber bullying. Many social media platforms can use AI algorithm as inspector for detecting bullying, automatically removing content that promotes any kind of bullying and harassment to keep the community safe.

4.1 Limitations

Firstly, the interpretation of ML algorithms can be influenced by accessibility of data and possible bias in training data. Secondly extant ML algorithms struggle with sarcasm and contextual nuances in language. Lastly, the validity of espials needs additional probing. Despite these limitations, this research underscores the significant potential of ML as a tool for combating cyberbullying.

4.2 Future Direction

By familiarizing with these limitations and nurturing reliable growth, ML algorithms can be further ameliorated to create a healthy online atmosphere for each and everyone. This

research furthers future investigations and the growth of firm ML solutions to adequately handle the roaring affair of cyberbullying.

References

Akhter, A., Acharjee, U. K., Talukder, M. A., Islam, M. M., & Uddin, M. A. (2023). A robust hybrid machine learning model for Bengali cyber bullying detection in social media. *Natural Language Processing Journal*, 4, 100027. https://doi.org/10.1016/j.nlp.2023.100027

Alduailaj, A. M., & Belghith, A. (2023). Detecting Arabic Cyberbullying Tweets Using Machine Learning. *Machine Learning and Knowledge Extraction*, 5(1), 29–42. https://doi.org/10.3390/make5010003

AlHarbi, B. Y., AlHarbi, M. S., AlZahrani, N. J., Alsheail, M. M., & Ibrahim, D. M. (2020). Using Machine Learning Algorithms for Automatic Cyber Bullying Detection in Arabic Social Media. *Journal of Information Technology Management*, (Online First). https://doi.org/10.22059/jitm.2020.75796

Ali, A., & Syed, A. M. (2020). Cyberbullying Detection Using Machine Learning.

Allen, S. (2019). Social media's growing impact on our lives. Retrieved March 10, 2024, from https://www.apa.org/members/content/social-media-research

Almomani, A., Nahar, K., Alauthman, M., Al-Betar, M. A., Yaseen, Q., & Gupta, B. B. (2024). Image cyberbullying detection and recognition using transfer deep machine learning. *International Journal of Cognitive Computing in Engineering*, 5, 14–26. https://doi.org/10.1016/j.ijcce.2023.11.002

Alqahtani, A. F., & Ilyas, M. (2024). An Ensemble-Based Multi-Classification Machine Learning Classifiers Approach to Detect Multiple Classes of Cyberbullying. *Machine Learning and Knowledge Extraction*, 6(1), 156–170. https://doi.org/10.3390/make6010009

Bergman, A. M. (2024). Effects of Cyberbullying. Retrieved March 10, 2024, from https://socialmediavictims.org/cyberbullying/effects/

Chen, S., Wang, J., & He, K. (2024). Chinese Cyberbullying Detection Using XLNet and Deep Bi-LSTM Hybrid Model. *Information*, 15(2), 93. https://doi.org/10.3390/info15020093

Dewani, A., Memon, M. A., & Bhatti, S. (2021). Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for Roman Urdu data. *Journal of Big Data*, 8(1), 160. https://doi.org/10.1186/s40537-021-00550-7

Ejaz, N., Razi, F., & Choudhury, S. (2024). Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm. *Computers in Human Behavior*, 153, 108123. https://doi.org/10.1016/j.chb.2023.108123

Ismail, N., Losada, D. E., & Ahmad, R. (2024). A Test Dataset of Offensive Malay Language by a Cyberbullying Detection Model on Instagram Using Support Vector Machine [Series Title: Communications in Computer and Information Science]. In N. H. Zakaria, N. S. Mansor, H. Husni & F. Mohammed (Eds.), *Computing and Informatics* (pp. 182–192, Vol. 2001). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-9589-9_14

Jadhav, R., Agarwal, N., Shevate, S., Sawakare, C., Parakh, P., & Khandare, S. (2023). Cyber Bullying and Toxicity Detection Using Machine Learning. *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, 66–73. https://doi.org/10.1109/ICPCSN58827.2023.00017

James, I. I., & Osubor, V. I. (2022). Hostile social media harassment: A machine learning framework for filtering anti-female jokes. *Nigerian Journal of Technology*, 41(2), 311–317. https://doi.org/10.4314/njt.v41i2.13

Kemp, S. (2024, January). Digital 2024: Global Overview Report. Retrieved March 10, 2024, from https://datareportal.com/reports/digital-2024-global-overview-report

Mahmud, T., Ptaszynski, M., & Masui, F. (2023). Automatic Vulgar Word Extraction Method with Application to Vulgar Remark Detection in Chittagonian Dialect of Bangla. *Applied Sciences*, 13(21), 11875. https://doi.org/10.3390/app132111875

Muneer, A., & Fati, S. M. (2020). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet*, 12(11), 187. https://doi.org/10.3390/fi12110187

Ostic, D., Qalati, S. A., Barbosa, B., Shah, S. M. M., Galvan Vela, E., Herzallah, A. M., & Liu, F. (2021). Effects of Social Media Use on Psychological Well-Being: A Mediated Model. *Frontiers in Psychology*, 12, 678766. https://doi.org/10.3389/fpsyg.2021.678766

Paul, C., Sahoo, D., & Bora, P. (2020). Aggression In Social Media: Detection Using Machine Learning Algorithms. 9(04).

Pericherla, S., & Ilavarasan, E. (2024). Overcoming the Challenge of Cyberbullying Detection inImages: A Deep Learning Approach with Image Captioningand OCR Integration. *International Journal of Computing and Digital Systems*, 15(1), 393–401. https://doi.org/10.12785/ijcds/150130

Raj, M., Singh, S., Solanki, K., & Selvanambi, R. (2022). An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques. *SN Computer Science*, 3(5), 401. https://doi.org/10.1007/s42979-022-01308-5

Singla, S., Lal, R., Sharma, K., Solanki, A., & Kumar, J. (2023). Machine Learning Techniques to Detect Cyber-Bullying. *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, 639–643. https://doi.org/10.1109/ICIRCA57980.2023.10220908

Sultan, D., Toktarova, A., Zhumadillayeva, A., Aldeshov, S., Mussiraliyeva, S., Beissenova, G., Tursynbayev, A., Baenova, G., & Imanbayeva, A. (2023). Cyberbullying-related

Hate Speech Detection Using Shallow-to-deep Learning. *Computers, Materials & Continua*, 74(1), 2115–2131. https://doi.org/10.32604/cmc.2023.032993

Tapsoba, W. C., Bassole, D., Kafando, R., Kabore, A. K., Sabané, A., & Bissyandé, T. F. (2024). Cyber Threat's detection using Machine Learning Algorithms.

Xingyi, G., & Adnan, H. M. (2024). Potential cyberbullying detection in social media platforms based on a multi-task learning framework. *International Journal of Data and Network Science*, 8(1), 25–34. https://doi.org/10.5267/j.ijdns.2023.10.021

Yuan, L., Wang, T., Ferraro, G., Suominen, H., & Rizoiu, M.-A. (2023). Transfer learning for hate speech detection in social media. *Journal of Computational Social Science*, 6(2), 1081–1101. https://doi.org/10.1007/s42001-023-00224-9