Journal of Information Technology and Digital World (ISSN: 2582-418X)
www.irojournals.com/itdw/

Cyberbullying Detection using Machine Learning Models

Kanitha T[1], Dhanya K.R.[2], Karpagam C.[3]

[1,2]Student, [3]Assistant professor, Computer Science with Data Analytics, Dr.N.G.P. Arts and Science College, Bharathiar university, Coimbatore, India

E-mail: [1]kanithathangavel33@gmail.com, [2]dhanyafab@gmail.com, [3]karpagam@drngpasc.ac.in

Abstract

Cyberbullying is a significant and increasing problem in online communities, and the detection system should also be effective in addressing it. The research presents an in-depth comparison of image classification systems such as Logistic Regression, Naive Bayes, XGBoost, Decision Tree, and Random Forest in the detection of cyberbullying. The evaluation of the five machine learning algorithms with respect to: Logistic Regression, Naive Bayes, XGBoost, Decision Tree, and Random Forest, will be within the framework of large-scale dataset collection about cyberbullying. This will be done based on the evaluation of the metadata file using accuracy, precision, recall, and F1 score, which represent the overall performance level. The results presented help determine the weaknesses and strengths of the individual algorithms and narrow the search for the right approach to cyberbullying detection. Moreover, best-performing algorithms were integrated into a Stream -lit- based front end for real-time prediction and display of the capabilities of the model. This study contributes significantly to the research on the development of new machine-learning solutions for cyberbullying detection and provides a solid evaluation of various classification strategies that are ultimately well-suited for effective detection systems in the future.

Keywords: Cyberbullying Detection, Classification Strategies, Machine Learning, Real-time Prediction.

DOI: https://doi.org/10.36548/jitdw.2024.4.004
Received: 18-12-2024. received in revised form: 04-01-2025, accepted: 20-01-2025, published: 31-01-2025

1. Introduction

Cyberbullying, which involves using digital media to humiliate, harass, or harm people, has become a major social issue with the advent of social media and messaging applications since the early 2000s. Unlike traditional bullying, which usually occurs face-to-face, cyberbullying can happen at any time and from anywhere due to the constant presence of the internet. This omnipresence makes it exceedingly difficult for victims to find refuge or respite from their aggressors [1,2].

In response to this alarming trend, machine learning presents a promising avenue for detecting and mitigating harmful online behavior. By analyzing vast amounts of data sourced from social media the machine-learning algorithms can identify the patterns and signals that indicate cyberbullying, such as offensive language, threats, or repeated insults. Most importantly, these algorithms are self-improving and refining over time as they process more and more data to eventually identify the harmful content more quickly and efficiently.

This technological advancement is very important for improving a safer online environment and for effectively tackling the widespread issue of cyberbullying. It aims to evaluate five different machine learning methods, namely Logistic Regression, Naive Bayes, XGBOOST, Decision Tree, and Random Forest to determine the most efficient methods that can be used in cyberbullying detection. Using a significant amount of data, this research aims to identify which of the used algorithms can obtain the highest degree of accuracy and efficiency in detecting cyberbullying.

Finally, the work demonstrates how an algorithm from this experiment can be incorporated into an application that deals with real-time data using Stream lit-based utility, demonstrating its effectiveness in quickly spotting and reacting to cyberbullying incidents as they occur. This shows not only the practical applications of the research but also the possibility of creating better systems focused on preventing and combating cyberbullying in any online environment [3,4]. Eventually, the information that will be obtained from this research will hopefully add value to developing machine learning-based solutions for detecting cyberbullying and further pave the way for intervention and protection to be more effectively taken in a digital community[5].

2. Related Work

Cyberbullying detection has evolved significantly, using both text-based and multimodal approaches. Text-based detection methods started with traditional machine learning models like Support Vector Machines (SVM), Naive Bayes, and Decision Trees, which analyze textual features to identify harmful content. More recently, deep learning techniques such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks have been used for a deeper understanding of language and context. While traditional models require less data and work well with feature engineering, deep learning models often achieve better accuracy but need large datasets and high computational power [1-4]. Beyond text, multimodal detection expands the scope by incorporating images and videos, using CNNs alongside text-based models to improve cyberbullying identification. When it comes to classifying cyberbullying, some systems rely on rule-based methods, such as keyword searches and heuristics, for cyberbullying content. Others use ensemble learning, like Random Forests, which combines multiple classifiers for better accuracy [5,6]. While rule-based approaches work for explicit cases, they struggle with context, whereas ensemble methods offer more reliability by considering diverse features [7,8]. Sentiment analysis further enhances detection by assessing the emotional tone behind messages, using tools like VADER, TextBlob, and emotion lexicons. While pre-trained sentiment models are useful, they sometimes miss the context, whereas supervised learning models can be fine-tuned for specific datasets but require extensive labelling [9]. Finally, intervention strategies help mitigate cyberbullying through automated moderation tools that analyse and identify the content in real time, often using RNNs for tracking patterns over time. Predictive models, based on historical data, attempt to identify potential future bullying incidents before they happen. However, real-time interventions can sometimes misinterpret content, leading to unnecessary censorship, while predictive models, though valuable for understanding long-term trends, may not provide immediate protection [10-14].

3. Proposed Work

A. Data Collection

This dataset is a collection of data from various sources related to the automatic detection of cyberbullying. The data originates from different social media platforms, including Kaggle, Twitter, Wikipedia Talk pages, and YouTube. It consists of text data labelled as either 'bullying' or 'not bullying,' and encompasses various types of cyberbullying, such as hate speech, aggression, insults, and toxicity. Table 1 shows a sample of the twitter_sexism_parsed_dataset [12] collected from Kaggle Cyberbullying Dataset. This dataset contains 14,881 unique values and uses binary labels '0' and '1' to represent non-cyberbullying and cyberbullying, respectively.

Table 1. Sample Dataset

index	id	Text	Annotation	oh_label
ID	ID	Text	Category	Category
14,881 unique values	14,881 unique values	14,881 unique values	none 77.3% sexism 22.7% blank 0%	0 77.3% 1 22.7% blank 0%
5.35198627292254E+017	5.35198627292254E+017	RT @BeepsS: @senna1 @BeepsS: I'm not sexist but fuc	sexism	1
5.75984924030714E+017	5.75984924030714E+017	There's some very hate able teams this year #MKR	none	0
5.7233536016588E+017	5.7233536016588E+017	RT @The_Eccles: "Everyone underestimated us" We sti	none	0
5.72337925708374E+017	5.72337925708374E+017	RT @NOTLukeDarcy: did @Channel7 or #MKR actually	none	0
4.43033024528011E+017	4.43033024528011E+017	No, you don't. @Shut_Up_Jeff: I thought of a really funn	sexism	1
5.69577286308987E+017	5.69577286308987E+017	RT @Wateronatrain: @MT8_9 You might like this http://t	sexism	1
5.75951008863429E+017	5.75951008863429E+017	RT @kholly265: I bet the campers vote strategically...at	none	0
5.73948678966108E+017	5.73948678966108E+017	@EvvyKube it is absurd how much of my amazon wish l	none	0
5.7233188575119E+017	5.7233188575119E+017	RT @DanielleVLee: Colin is obviously malnourished fror	none	0
5.69655750961668E+017	5.69655750961668E+017	@NewsCoverUp @RJennromao @GBabeuf @DavidJo5	none	0
5.68436168649343E+017	5.68436168649343E+017	RT @MetalBarbieDoll: But yea, apparently #GamerGate	sexism	1
5.7559934997708E+017	5.7559934997708E+017	"@Sam_1985: Notice we didn't see Kat and Andre in an	none	0
5.61984177701421E+017	5.61984177701421E+017	RT @g56yu: @PierceCotwa is now on twitter. If u care a	none	0
4.2696640533903E+017	4.2696640533903E+017	:D @nkrause11 Dudes who go to culinary school: #why	sexism	1

B. Data Cleaning

The dataset was in .CSV format. Because the fields were straightforward, the original fields in the annotation attributes were removed and replaced with label values to make the next steps easier.

C. Data Preprocessing

The pre-processing phase is essential for preparing the dataset for machine learning analysis and involves several detailed steps:

1. **Word Tokenization**: This step involves breaking down the text into individual words or tokens. Tokenization is essential as it converts the raw text into a structured format

that can be processed further. Each sentence or paragraph is split into a list of words, which serves as the basic unit of analysis for subsequent steps.

2. **Stop Words Filtering**: Stop words are common words that carry little meaning on their own and are often filtered out during pre-processing.
    a. Using NLTK's stopwords.words('english'), we removed these words (e.g., "the", "a", "an") from the dataset. This step is important because stop words do not contribute significantly to the meaning of the text and can distort the analysis by adding noise.
3. **Punctuation Removal**: Punctuation marks such as commas, periods, and exclamation points were removed from the text. This was achieved by retaining only the characters that are not punctuation, as identified using string.punctuation. Removing punctuation helps in focusing on the core textual content and prevents punctuation from affecting the analysis.
4. **Stemming**: Stemming involves reducing words to their base or root form. For instance, words like "connection", "connected", and "connecting" are reduced to the root word "connect" using NLTK's PorterStemmer. This normalization helps in grouping different forms of the same word, thus improving the consistency of the text analysis.
5. **Digit Removal**: Numeric content was filtered out from the text. Since numbers do not contribute to the context of cyberbullying detection, removing them helps in focusing solely on the textual content of the tweets.
6. **Feature Extraction**: The final pre-processing step involved extracting features using the Term Frequency-Inverse Document Frequency (TF-IDF) method. This technique balances word frequency in the document with its frequency across all documents, highlighting words that are significant to specific documents. This transformation prepares the text data for input into machine learning algorithms.

These pre-processing steps ensure that the dataset is clean, consistent, and ready for effective analysis and modelling in the cyberbullying detection study.

D. Description of Models

The proposed work compares the classification performance of five different machine learning models: Naïve Bayes, Logistic Regression, Random Forest Classifier, Decision Tree, and XGBoost. The classifiers were implemented in Python using the sklearn.naive_bayes, sklearn.linear_model, and sklearn.ensemble packages. The dataset was split into 70% for training and 30% for testing the models, respectively. Table 2 below shows the hyperparameter values used. Randomized search followed by grid search was used to optimize the hyperparameters.

Table 2. Hyperparameters and Values

Model	Hyperparameters	Values
Naïve Bayes	Alpha	1.0
	Fit Prior	True
Logistic Regression	C (inverse regularization strength)	1
	Solver	Liblinear
	Penalty	L2
Random Forest	No. of estimators	200, 500
	Maximum Depth	10, 20
	Minimum Samples Split	2,5
	Minimum Sample Leaf	1,2
	Bootstrap	True
Decision Tree	Maximum Depth	10,20
	Minimum Samples Split	2,5
	Minimum Sample Leaf	1,2
	Splitting Criterion	gini
XGBoost	Learning Rate	0.1

	Maximum Depth	6
	No. of estimators	500
	Subsample	0.8
	Gamma	0
	L1 regularization	1
	L2 regularization	0

4. Experiment and Results

For our analysis of supervised learning techniques, we evaluated Naive Bayes, Logistic Regression, and Decision Tree as standard methods, and XGBoost and Random Forest Classifiers as ensemble methods. We observed that the XGboost performed the best across all metrics, while the Naive Bayes classifier was the least effective. XGBoost and Random Forest perform the best, achieving an accuracy of 0.90 and 0.94, respectively. Logistic Regression also shows strong performance with an accuracy of 0.84, making it a good choice for a balance between performance and interpretability. Naïve Bayes struggles with precision and recall, making it less ideal for this dataset. Meanwhile, the Decision Tree model performs slightly worse than Random Forest, as expected, due to its lack of ensemble learning, which limits its ability to capture complex patterns compared to Random Forest. The results are depicted in Table 3. The Figure .1 below shows the graphical representation of the results observed

Table 3. Performance of Machine Learning Methods

Metrics/Models	Naïve Bayes	Logistic Regression	Random Forest	Decision Tree	XGBoost
Accuracy	0.64	0.84	0.90	0. 88	0.94
Precision	0.74	0.85	0.91	0.89	0.95
Recall	0.64	0.84	0.90	0.87	0.94
F1-Score	0.61	0.84	0.90	0.87	0.94
ROC-AUC	0.69	0.86	0.92	0.89	0.95

Confusion Matrix	960 1482 45 1514	1950 520 220 1311	2130 280 80 1490	2020 390 150 1400	2150 260 70 1500

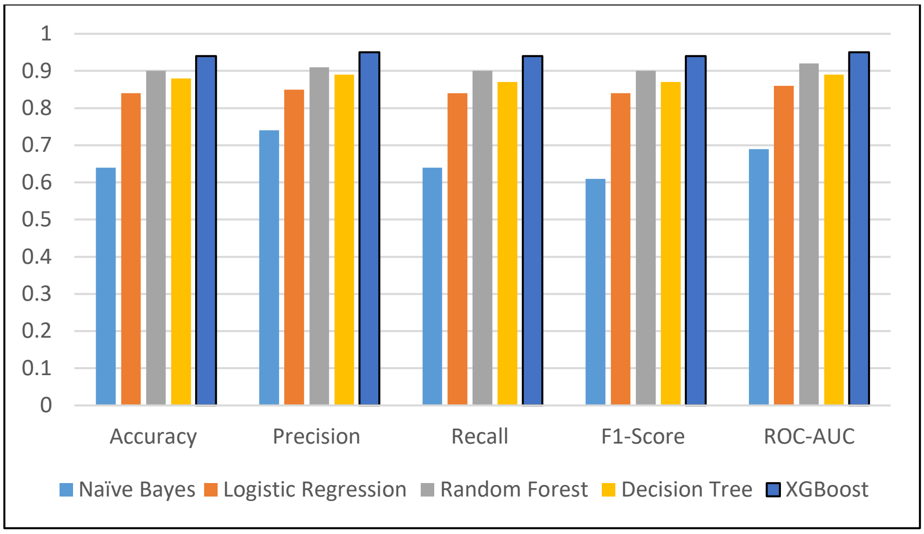
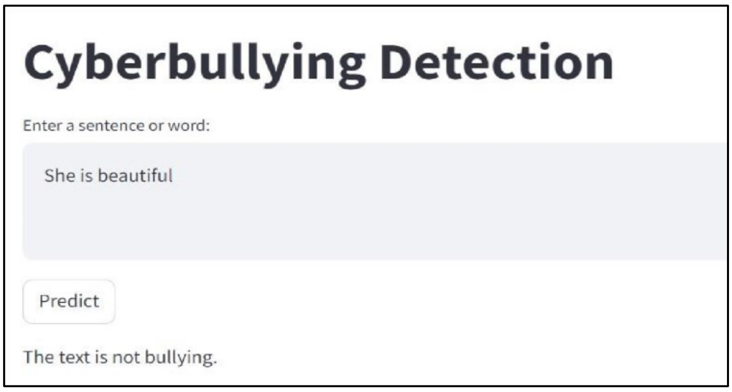
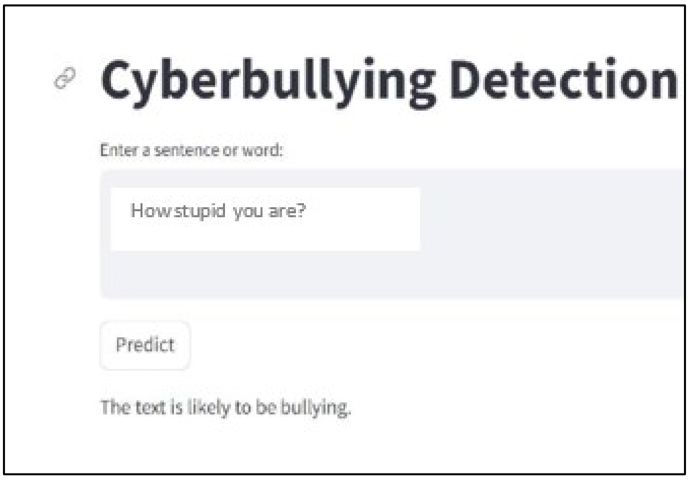


Figure 1. Performance Comparison of Machine Learning Models

Streamlit an open-source framework was used for creating interactive web applications with Python. Streamlit can be used to develop a web application for identifying cyberbullying using the XGBoost model, which was opted due to its superior classification accuracy among the five models compared. The trained XGBoost model is integrated into the app, where Streamlit provides an interface for users to input text, and the model predicts if it's cyberbullying, ensuring efficient and user-friendly real-time classification Figure 2 depicts the application results on real-time classification.



(a)



(b)

Figure 2. (a) (b) Detection of Cyberbullying using Web Application

5. Conclusion

The study conducted a comparative analysis of multiple machine learning algorithms for the detection of cyberbullying, revealing that the Random Forest classifier achieved the highest accuracy, reaching 92%. Notably, ensemble methods consistently outperformed traditional algorithms, with the Naive Bayes classifier exhibiting the lowest accuracy at 61%. To enhance the practical applicability of our findings, we integrated the top-performing Random Forest model with a Streamlit-based frontend, facilitating real-time predictions. This integration supports the development of effective solutions for cyberbullying detection. Future research could explore advanced models like SVM and MLP for improved accuracy, use semi-supervised learning to address limited labeled data, and incorporate multimodal data (text, images, videos) for more comprehensive detection. Expanding data diversity across social media platforms, languages, and cultures would enhance generalizability and system effectiveness.

References

[1] Al-Garadi, Mohammed Ali, Mohammad Rashid Hussain, Nawsher Khan, Ghulam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak, and Abdullah Gani. "Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges." IEEE Access 7 (2019): 70701-70718.

[2] Raj, Chahat, Ayush Agarwal, Gnana Bharathy, Bhuva Narayan, and Mukesh Prasad. "Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques." Electronics 10, no. 22 (2021): 2810.

[3] Ahmed, Md Tofael, Maqsudur Rahman, Shafayet Nur, Abu Zafor Muhammad Touhidul Islam, and Dipankar Das. "Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts." TELKOMNIKA (Telecommunication Computing Electronics and Control) 20, no. 1 (2021): 89-97.

[4] Kumar, Raju, and Aruna Bhat. "A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media." International Journal of Information Security 21, no. 6 (2022): 1409-1431.

[5] Wang, Shuai, Abdul Samad Shibghatullah, Thirupattur Javid Iqbal, and Kay Hooi Keoy. "A review of multimodal-based emotion recognition techniques for cyberbullying detection in online social media platforms." Neural Computing and Applications 36, no. 35 (2024): 21923-21956.

[6] Dinakar, Karthik, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. "Common sense reasoning for detection, prevention, and mitigation of cyberbullying." ACM Transactions on Interactive Intelligent Systems (TiiS) 2, no. 3 (2012): 1-30.

[7] Silva, Y. N., Hall, D. L., & Rich, C. (2018). BullyBlocker: Toward an interdisciplinary approach to identify cyberbullying. Social Network Analysis and Mining, 8. doi:10.1007/s13278-018-0496-z

[8] Zych, Izabela, Rosario Ortega-Ruiz, and Rosario Del Rey. "Systematic review of theoretical studies on bullying and cyberbullying: Facts, knowledge, prevention, and intervention." Aggression and violent behavior 23 (2015): 1-21.

[9] Naf'an, Muhammad Zidny, Alhamda Adisoka Bimantara, Afiatari Larasati, Ezar Mega Risondang, and Novanda Alim Setya Nugraha. "Sentiment analysis of cyberbullying on instagram user comments." Journal of Data Science and Its Applications 2, no. 1 (2019): 38-48.

[10] Kaur, Manpreet, and Munish Saini. "Indian government initiatives on cyberbullying: A case study on cyberbullying in Indian higher education institutions." Education and Information Technologies 28, no. 1 (2023): 581-615.

[11] Alabdulwahab, Aljwharah, Mohd Anul Haq, and Mohammed Alshehri. "Cyberbullying Detection using Machine Learning and Deep Learning." International Journal of Advanced Computer Science and Applications 14, no. 10 (2023).

[12] https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset

[13] Robert, Abil. Combating Cyberbullying with Machine Learning and Deep Learning. No. 13035. EasyChair, 2024.

[14] Agrawal, T., B.Tech, & Chakravarthy, D.V. (2022). Cyberbullying Detection and Hate Speech Identification using Machine Learning Techniques. 2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS), 182-187.