Research Article/Review Article

Journal of Computing, Science &Technology. Vol. 2, 2024

Journal of Computing, Science &Technology

https://focjournal.unidel.edu.ng/
editorjcst@unidel.edu.ng

ISSN: 3043-6818 Print

https://focjournal.unidel.edu.ng

Hybrid Predictive Modeling for Early Cyberbullying Detection Using Deep Learning and Gradient Boosting

Ojulowo Lekan Omolere[1], Oluwadare Samuel Adebayo[2], Ayeni Olaniyi Abiodun[3]
[1]Department of Computer Science, School of Computing, The Federal University of Technology, Akure, Ondo State, Nigeria
[2]Department of Computer Science, School of Computing, The Federal University of Technology, Akure, Ondo State, Nigeria
[3]Department of Cyber Security, School of Computing, The Federal University of Technology, Akure, Ondo State, Nigeria
ojulowolo@futa.edu.ng[1], saoluwadare@futa.edu.ng[2], oaayeni@futa.edu.ng [3]

Corresponding Author's Email: ojulowolo@futa.edu.ng

ABSTRACT

Article Info

Date Received: 29-10-2024
Date Accepted: 15-12-2024

Cyber bullying or cyber harassment is a form of bullying or harassment using electronic means. It involves posting rumors, sexual remarks, pejorative labels (hate speech), etc. It does not require physical power nor strength of numbers. Victims of cyberbullying may experience depression, ill-health, low self-esteem, increased suicidal ideation, emotional disorder, and negative emotional responses such as being scared, frustrated and angry. Social Media and the Internet have opened up new forms of both empowerment and oppression. Meaningful engagement has transformed into a detrimental avenue where individuals are often vulnerable targets of online ridiculing and cyberbullying. The rise of cyberbullying on social media necessitates efficient and accurate detection systems to safeguard users. In this paper, a hybrid approach that combines the strengths of deep learning and gradient boosting algorithms is used to develop a predictive model for early cyberbullying detection. Specifically, a Long Short-Term Memory (LSTM) network is used to extract temporal and contextual sequences from text data, and a Gradient Boosting Machine (GBM) for early prediction of cyberbullying. The LSTM cell state runs through the entire sequence allowing the model to maintain long-term dependencies and carrying essential information throughout the processing of the sequence regulated by structures called gates. The gradient boosting machine works by building an ensemble of weak learners in a sequential manner, where each new learner tries to correct the errors made by the previous ones. This leads to a strong predictive model that can handle complex relationships in data. The predictive model for early detection of cyberbullying developed from this research has proven to be an efficient way of providing suitable model for predicting, detecting and mitigating cyberbullying threats and potential risks. Experimental results on benchmark datasets from twitter, reddit and kaggle indicated that the model achieves higher precision and recall than standalone models.

Keywords:
Cyberbullying, Hybrid Model, LSTM, Gradient Boosting, Predictive Modeling.

1.0 INTRODUCTION

As social media usage grows, cyberbullying has become a significant issue, particularly among adolescents. Detecting and intervening early in bullying incidents can reduce their negative impacts. Although many machine learning techniques have been applied to this problem, there is a need for a predictive model that combine both the temporal aspects of online communication and robust classification algorithms for early detection of cyber bullying to prevent and safeguard online users from becoming victims of cyberbullying.

The use of technology to harass, threaten, embarrass, or target another person is known as cyberbullying [5-6]. Online threats, hostile or impolite texts, tweets, postings, or messages are its defining characteristics. Every time the victim checks their computer or gadget, they may experience constant torture. Because the bully does not have to physically face the victim, it is easier to commit than other forms of bullying.

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to model sequential data and capture long-term dependencies and relationships in sequence data. Gradient Boosting is a machine learning algorithm used for cyberbullying detection to build a better detection model by building an ensemble of weak learners (typically decision trees).

This paper introduces a hybrid model combining deep learning (LSTM) for feature extraction and Gradient Boosting for classification with the aim to detect cyberbullying as early as possible and put into consideration context dependent challenge of cyberbullying.

Qianjia *et al* (2014), explored the value of social information in detecting cyberbullying above and beyond the signals available in textual content. It identified both social and textual features to create a composite model for detecting cyberbullying. It does not consider the context of the contents and signs for early identification of the signals. **Li (2016),** developed a system that can predict bullying attacks on images posted on social media sites using Support Vector Machine (SVM) and Binary Decision tree to learn the pattern of bullied images and use the pattern to detect the bulliability of other images. It is does not make provision to identify images that contains bully contents as text, and not visually bullied.

Lu and Mazumder (2018), used gradient boosting algorithm by selecting a random subset of weak-learners and then find the best candidate among them using a greedy strategy in the prediction space to predict cyberbullying contents. The greedy strategy does not make provision for underused words and new words that are bully words.

Saxena (2019), developed a hybrid deep learning model for bullying content prediction, where the content c ε {text, image, info-graphic}. The hybrid architecture consists of a convolution neural network (CNN) for predicting the textual bullying content and a support vector machine (SVM) classifier for predicting the visual bullying content. Consideration for context-dependent cyberbullying content pose a challenge.

Alam et al (2021), Single Level Ensemble Model (SLE) and Double Level Ensemble Model (DLE) were used to identify and classify social media posts as 'offensive' and 'non-offensive' contents using ensemble-based voting model. It is limited in accommodating contents outside the labelled contents and contexts.

Raj et al (2022), developed a deep learning classification model that classified collected data and the classifier was used to detect cyberbullying content by eliminating any gray possibilities. The classifier was not enhanced to support context dependent contents.

Amran & Hassan (2023), explored and measured the use of auto-coding links in ATLAS.ti23 Qualitative analysis was used to identify two document groups in thematic categories to explore the conceptual perspective of cyberbullying prevention using ATLAS.ti23. The identified document groups are based on contextual words within the scope of the ATLAS.ti23

2.0 METHODOLOGY

The proposed hybrid model comprises two stages. In the first stage, an LSTM network is trained to capture the sequential and contextual information in text data putting into consideration context dependent data. The LSTM processes online posts from twitter, reddit and kaggle to produce feature representations. These features are then fed into a Gradient Boosting classifier, XGBoost, which is known for its high accuracy in structured data classification. The hybrid approach combines the LSTM's strength in capturing long-term dependencies with Gradient Boosting's ability to perform fine-grained classification.

The model was evaluated on multiple social media posts using twitter, reddit and kaggle, focusing on early detection performance.

Model Design Flow

Figure 1shows the flow of the design model.

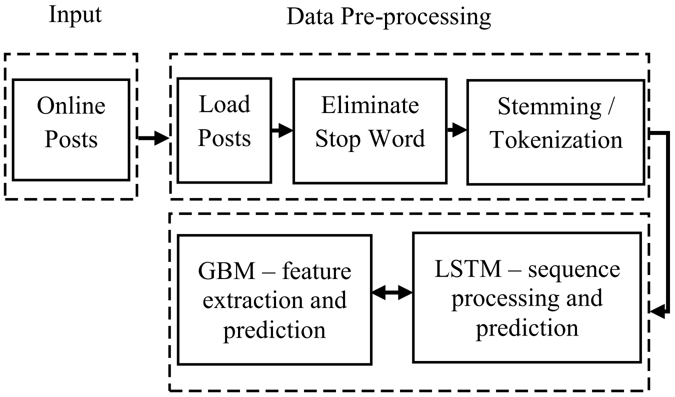


Fig 1: Model Design

2.1 Input

The input to the model consist of online posts obtained from online repository website (kaggle) for historical data, social media websites (twitter and reddit), and simulated dataset. The datasets varied across platforms and were stored as csv file.

2.2 Data Pre-processing

The dataset were loaded into the data pre-processing, which format the dataset by removing high frequency words such as it, as, the, I, etc., that have less sematic weight. The words were reduced to their base/stem form through stemming process and tokenization.

2.3 Hybrid Model for Data Processing

The processed data from the data pre-processing module were loaded into the data processing module, while the hybrid model is applied using Long Short-Term Memory (LSTM) sequential algorithm and Gradient Boosting Machine (GBM) feature extraction.

The LSTM cell state runs through the entire sequence allowing the model to maintain long-term dependencies and carrying essential information throughout the processing of the sequence. It is regulated by structures called gates.

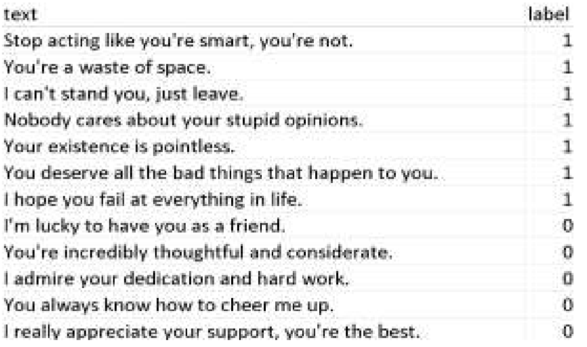
Gradient Boosting Machine works by building an ensemble of weak learners (typically decision trees), where each new tree tries to correct the errors made by the previous ones. This leads to a strong predictive model that handled complex relationships and context dependence in the data.

3.0 RESULT

The implementation of this model includes data preprocessing, model training, evaluation, and a hybrid ensemble technique. The model was implemented with Python programming language with some of the libraries and packages such as scikit-learn, numpy, and pandas, in developing the hybrid model for early cyberbullying detection using both LSTM (Long Short-Term Memory) and Gradient Boosting Machine (GBM) algorithms.

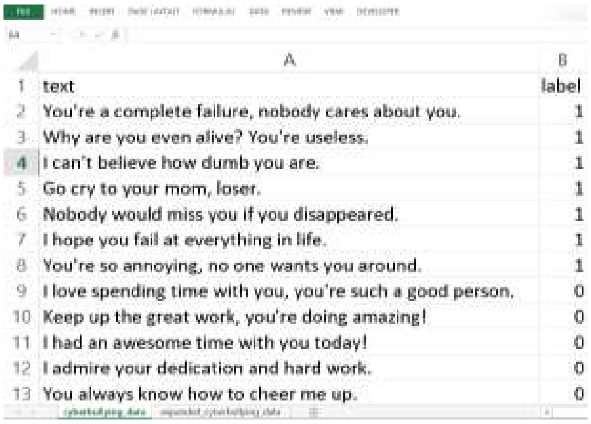
The dataset were obtained from online data repository website (kaggle, twitter, reddit), and were also simulated using the python library package known as pandas and saved as a character separated value (.csv) file.

Figure 2 and Figure 3 show the different dataset as a csv file, categorized as either bullying (1) or non-bullying (0).



text	label
Stop acting like you're smart, you're not.	1
You're a waste of space.	1
I can't stand you, just leave.	1
Nobody cares about your stupid opinions.	1
Your existence is pointless.	1
You deserve all the bad things that happen to you.	1
I hope you fail at everything in life.	1
I'm lucky to have you as a friend.	0
You're incredibly thoughtful and considerate.	0
I admire your dedication and hard work.	0
You always know how to cheer me up.	0
I really appreciate your support, you're the best.	0

Figure 2: Simulated Dataset as CSV



	A	B
1	text	label
2	You're a complete failure, nobody cares about you.	1
3	Why are you even alive? You're useless.	1
4	I can't believe how dumb you are.	1
5	Go cry to your mom, loser.	1
6	Nobody would miss you if you disappeared.	1
7	I hope you fail at everything in life.	1
8	You're so annoying, no one wants you around.	1
9	I love spending time with you, you're such a good person.	0
10	Keep up the great work, you're doing amazing!	0
11	I had an awesome time with you today!	0
12	I admire your dedication and hard work.	0
13	You always know how to cheer me up.	0

Figure 3: Online Dataset as CSV

3.1 Evaluation Metrics

It is a sensitive issue to identify non-bullying case as a bullying case (false positive) and to identify bullying content as normal (false negative). Thus, accuracy, precision, recall and the F1-score were considered for the performance evaluation metrics.

Precision: The total number of correctly predicted true bullying contents out of retrieved bullying contents.

P = TP / (TP + FP), where: P = Precision, TP = True Positive, FP = False Positive

Recall: Number of correctly predicted bullying cases from the total number of true bullying cases.

R = TP / (TP + FN), where: R = Recall, TP = True Positive, FN = False Negative

F1-score: The equally weighted harmonic mean of precision and recall. The model present the classifier performance on the average value.

F1-score = 2 / (1/R + 1/P), where: P = Precision, R = Recall

Gradient Boosting Machine (GBM) Model Evaluation

Table 1: Performance Evaluation of the GBM Model

Metrics	**GBM**
Accuracy	0.85
F1-Score	0.79
Precision	0.82
Recall	0.76

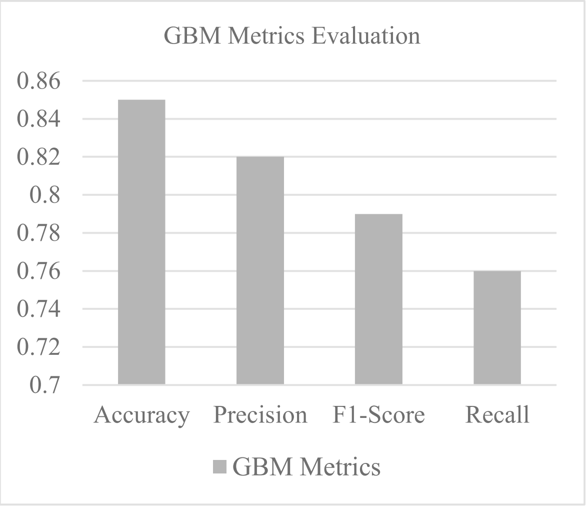


Chart 1: Evaluation of the GBM Model Performance

Support Vector Machine (SVM) Model Evaluation

Table 2: Performance Evaluation of the SVM Model

Metrics	**SVM**
Accuracy	0.86
F1-Score	0.80
Precision	0.84
Recall	0.77

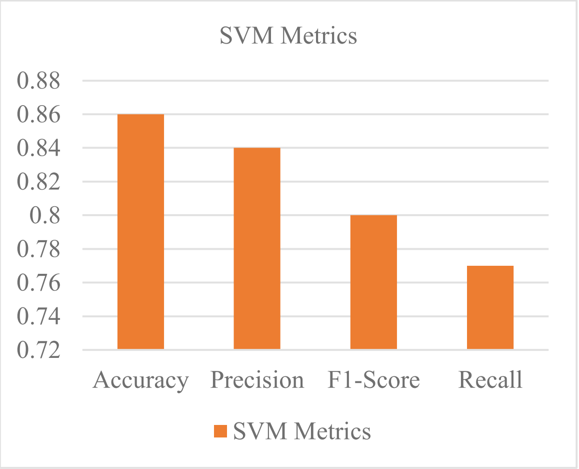


Chart 2: Evaluation of the SVM Model Performance

Long Short-Term Memory (LSTM) Model Evaluation

Table 3: Performance Evaluation of the LSTM Model

Metrics	**LSTM**	**HYBRID Model**
Accuracy	0.87	0.89
F1-Score	0.81	0.84
Precision	0.85	0.86
Recall	0.78	0.82

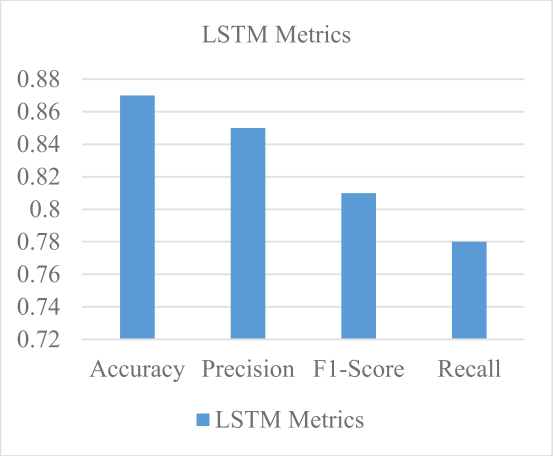


Chart 3: Evaluation of the LSTM Model Performance

Hybrid Predictive Model Evaluation

Table 4: Performance Evaluation of the HYBRID Model

Metrics	**HYBRID Model**
Accuracy	0.89
F1-Score	0.84
Precision	0.86
Recall	0.82

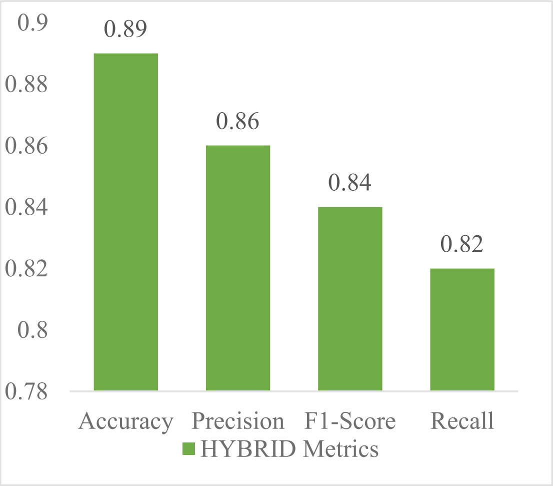


Chart 4: Evaluation of the HYBRID Model Performance

Table 5: Performance Evaluation of the Hybrid Model against Individual Models

Metrics	**GBM**	**SVM**	**LSTM**	**HYBRID Model**
Accuracy	0.85	0.86	0.87	0.89
F1-Score	0.79	0.80	0.81	0.84
Precision	0.82	0.84	0.85	0.86
Recall	0.76	0.77	0.78	0.82

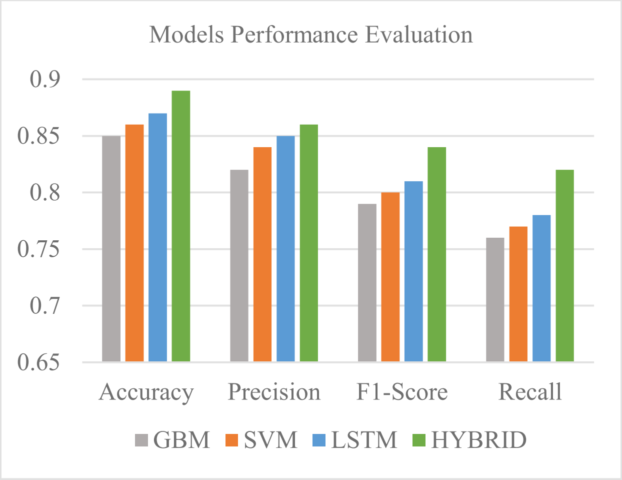


Chart 5: Evaluation of the Hybrid Model Performance against Individual Models

The hybrid model performs better than the individual models.

Result

The hybrid model shows superior performance compared to individual models. The use of Gradient Boosting significantly improves the prediction accuracy, particularly in handling context dependent data and detecting subtle bullying behaviors, making it more reliable for early detection of cyberbullying.

4.0 Discussion

To address the interpretability challenges of a hybrid LSTM and Gradient Boosting Model (GBM), **LIME (Local Interpretable Model-Agnostic Explanations) is used** to explain the predictions made by the Gradient Boosting component. LIME generates a local surrogate model around the specific prediction.

For the input

LIME output:

- Positive class (Bullying):
    - "loser": +0.64
    - "nobody likes": +0.58
- Negative class (Non-Bullying):
    - "are": -0.05

This helps to highlight the words most influential in classifying the text as cyberbullying.

The combination of deep learning and gradient boosting techniques in a hybrid model offers a powerful tool for early cyberbullying detection. This approach balances the need for temporal and semantic understanding of conversations with strong performance, providing a promising direction for future research and practical applications in cyberbullying prevention.

This research has contributed to knowledge by establishing a hybrid model that helps in early detection of cyberbullying by identifying and mitigating cyberbullying threats, risks and damages.

The hybrid model for early detection of cyberbullying developed from this research has provided an efficient model for identifying and mitigating cyberbullying threats. It is recommended for researchers especially those in the computer science field and cyber security.

Future works could incorporate multiple and advance machine learning techniques in providing multi-modal model for classifying cyberbullying.

REFERENCES

[1]. Alam, Kazi & Bhowmik, Shovan & Prosun, Priyo. (2021). Cyberbullying Detection: An Ensemble Based Machine Learning Approach.

[2]. Amran, K., & Hassan, M. S. (2023). A Review on Cyberbullying Prevention in Social Media among Adolescents. *International Journal of Academic Research in Business and Social Sciences*, *13*(3), 174 – 189. Accessed May 10, 2023

[3]. Haihao Lu and Rahul Mazumder (2018). Randomized Gradient Boosting Machine. https://web.mit.edu/haihao/www/papers/RGBM.pdf. Accessed December 1, 2023.

[4]. Hao Li (2016). Image Analysis of Cyberbullying Using Machine Learning Techniques.

[5]. Help Guide. Cyberbullying: Dealing With Online Bullying. https://www.helpguide.org/articles/abuse/cyberbullying-dealing-online-bullies.htm. Accessed March, 28th 2023.

[6]. Lawrence Robinson and Jeanne Sega. Bullying and Cyberbullying. How to Deal with a Bully and Overcome Bullying. www.helpguide.org/articles/abuse/bullying.htm. Updated March 16th, 2023. Accessed March 28th, 2023.

[7]. Mudita Saxena (2019). Hybrid Deep Learning Model for Cyberbullying Detection on Social Multimodal Data.

[8]. Qianjia Huang, Vivek K. Singh and Pradeep K. Atrey (2014). Cyber Bullying Detection Using Social and Textual Analysis. Association for Computing Machinery (ACM).

[9]. Raj, M., Singh, S., Solanki, K. *et al* (2022). An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques. *SNCOMPUT. SCI.* 3, 401. https://doi.org/10.1007/s42979-022-01308-5. Accessed June 22, 2023.