ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1058

Machine Learning-Based Strategies for Detecting Cyberbullying in Online Chats

Victor Ojodomo Akoh[1]; Fati Oiza Ochepa
Department of Computer Science
Federal University Lokoja Kogi State, Nigeria

Abstract:- This study employed the stacking of three machine learning techniques: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Logistic Regression algorithms to develop a model for detecting cyberbullying using a post dataset acquired from the X Platform. The proposed model's task is to extract keywords from the post dataset and then classify them as either 1 ("cyberbullying word") or 0 ("not cyberbullying word"). The model generated an accuracy of 85.52%, and it was deployed using a simple Graphical User Interface (GUI) web application. This study recommends that the model be included on social media platforms to help reduce the growing use of cyberbullying phrases.

Keywords:- *Cyberbully, Machine Learning, Detection, Social Media.*

I. INTRODUCTION

The Internet's global accessibility has significantly changed our perception of the world. Social media (SM) is a derivative of the World Wide Web; social media usage is becoming increasingly popular, and it encompasses a variety of forms, including online news forums, gaming platforms, and dating apps, as well as social networking sites (e.g., Instagram, Facebook, X, etc.). People of various ages, origins, and social and economic classes have gradually incorporated social networking into their lives. Social media facilitates connections between people from all around the world. X, a social media platform for opinion transmission and image/video sharing, has surely become one of the most popular social networking platforms, allowing users to upload photographs and videos for other users to view and comment on.

Recently, there have been growing concerns about the usage of social media platforms (especially X) to disseminate opinions that may be broadly categorized as offensive. These offensive posts may manifest as hate speech, cyberbullying, and other similar forms of content. Hate speech is speech that diminishes or disparages an individual or a collective based on their origin, ethnicity, sexuality, gender identity, disability, religious beliefs, and political affiliation [1]. Cyberbullying refers to the act of using electronic communication mediums to harass and intimidate people by sending them malicious messages via platforms such as social media, instant messaging, or digital texts. An online bully is an individual who uses the internet, cell phones, or other technological devices to send harmful, shaming, threatening, tormenting, humiliating, or intimidating emails, as well as upload text or photographs on social media platforms, with the intention of causing harm to their victim [1].

Individuals have found social media platforms an easier alternative to communicate their thoughts, feelings, and emotions to their peers. The act of cyberbullying perpetrated by a person on a social media platform can have detrimental effects on the victim's physical and emotional well-being; in extreme cases, it can even result in suicidal thoughts, self-harm, and loss of life. Research reveals that cyberattacks primarily target teenagers and young adults. Owing to the large number of young people who are actively using social media platforms like X, cyberbullying has become a significant problem that has increasingly affected the online community [2]. An efficient approach to tackling this problem is to detect and encrypt the bullying messages prior to their delivery to the intended recipient. The purpose of this study is to enhance the current cyberbullying detection system through the utilization of the stacking ensemble technique.

II. RELATED WORK

Research by [3] on parameterized optimization neural network frames was the focus of the research work. It involved an algorithmic comparison of eleven categorization algorithms, out of which logistic regression yielded the best result. Bi-GRU and Bi-LSTM performed the best out of the neural networks utilized. The researchers' proposed shallow neural network outperformed the existing state-of-the-art techniques based on the accuracy and f1 score of 95% and 98%, respectively.

[4] used four deep learning models convolutional neural network (CNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BLSTM), and Contextual Long Short-Term Memory (CLSTM) and five machine learning models (Naive Bayes, Support Vector Machine, IBK, Logistic Regression, and JRip) for detecting abusive phrases in Urdu and Roman Urdu comments, respectively. Their results revealed that CNN had optimal performance with accuracies of 96.2% and 91.4% for one-layer and two-layer designs, respectively. They concluded based on the results that deep learning models outperform machine learning models and that deep learning models with one-layer designs generate more accurate results than two-layer designs.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1058

The research work of [5] involved using supervised learning to track out and stop cyberbullying on X platform. The model is based on both support vector machines (SVM), Naive Bayes, and the TFI-DF vectorizer for data mining. The researchers developed a methodology to protect people from online media threats, and the results showed that support vector machines outperformed Naive Bayes in identifying social media bullying content.

Data from two prominent sources of cyberbullying was utilized by [6]: hate speech posts gotten from X and personal assault comments originating from Wikipedia. The researcher created a model for detecting cyberbullying in text data using natural language processing and machine learning. The researcher adopted three feature extraction methods and four classifiers to determine the optimum technique. The developed model yielded an accuracy of over 90% for posts and 80% for Wikipedia data.

Embedded sentiment and lexicon characteristics were used by [7] in a supervised machine learning approach for the detection of cyberbullying on X platform and categorizing the degree of bullying into multi-class categories. Random Forest, Support Vector Machine, Naïve Bayes, Decision Tree and KNN were the machine learning techniques utilized in the extraction of features. The study findings indicated that the framework that was developed offered a feasible option for the identification of cyberbullying instances and assessing its severity level in online social networks, and that after comparing the results obtained from testing the baseline feature and proposed features on the different machine learning techniques, the proposed features are as important in detecting cyberbullying.

The researchers in [8] utilized supervised machine learning techniques to identify and address instances of cyberbullying in their study. Multiple classifiers were employed to train and detect instances of bullying behavior. The approach recommended by this study outperformed SVM on the cyberbullying dataset, achieving an accuracy of 92.8% compared to SVM's accuracy of 90.3%. Using the same dataset, a neural network (NN) demonstrated superior performance compared to other classifiers that performed similar tasks.

These studies utilized a range of machine learning and deep learning techniques, showcasing the effectiveness of different models and approaches in detecting cyberbullying. The identified research gaps include the need for ensemble techniques combining multiple models, handling imbalanced data, improving detection in different languages, providing granular categorization of cyberbullying severity, comparing effectiveness across social media platforms, enhancing real-time detection and deployment, and utilizing comprehensive evaluation metrics beyond accuracy and f1 scores. This research builds on these works by proposing a stacking ensemble technique combining SVM, KNN, and logistic regression to improve cyberbullying detection accuracy on X post datasets

III. METHODOLOGY

This research engaged the use of a stacking classifier made up of three machine learning algorithms: Support Vector Machine (SVM), Logistic Regression (LR), and K-Nearest Neighbor (KNN).

A. Stacking Classifier

The Stacking Classifier is a library in Scikit-Learn that combines two or more regression or classification models with the aim of improving performance. The structure comprises of two tiers of estimators. The first layer consists of all the baseline models, which predict the results of the test dataset. The second layer consists of a meta-classifier or regressor that generates new predictions by utilizing the predictions made by the baseline models as its input.

B. Support Vector Machine

A support vector machine technique is employed to depict different classes in a hyperplane within a multidimensional space. The SVM model generates the hyperplane in an iterative manner to minimize error. The objective of Support Vector Machines (SVM) is to partition datasets into distinct groups by identifying a hyperplane with the largest margin [9].

C. Logistic Regression

Logistic regression (LR) is an algorithm that uses the logistic function to create a distinct hyperplane between two datasets. The logistic regression algorithm employs the attributes (inputs) in order to generate a prediction that aligns with the likelihood of a suitable class for the given input [7].

D. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a fundamental technique employed in machine learning for the purpose of classification. Machine learning models employ various input variables to forecast output values. KNN is a rudimentary type of machine learning algorithm, mostly employed for classification purposes. Data points are classified based on the classification of their nearest neighbor, using a similarity metric with previously stored data points. [10].

Fig. 1. illustrates the division of the system development process into smaller, interrelated sub-activities to effectively accomplish the research's objective.

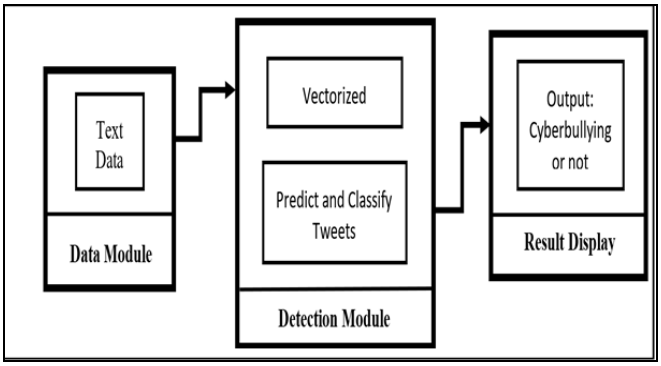

Fig 1: System Design of Cyberbullying Detection System

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1058

IV. RESULTS AND DISCUSSION

A. Result Presentation

The objective of this study is to enhance the current cyberbullying detection model. Our study employed logistic regression, KNN, SVC, and the stacking method with accuracy of 85.26%, 72.44%, 84.01%, and 85.52%. The stacking approach is the algorithm that achieved the highest performance, with an accuracy rate of 85.52%.

The model was developed using the X dataset, comprising 16,851 posts or texts, each with annotation features indicating none, sexism, or racism, and related label features represented by 1 or 0, denoting the presence of cyberbullying or its absence in the phrase, respectively. The label feature is slightly imbalanced because it comprises 5347 posts classified as cyberbullying and 11,501 posts classified as non-cyberbullying. For the model building, after pre-processing, the dataset was partitioned into 80% training data and 20% testing data and deployed on four algorithms: logistic regression, KNN, SVC, and the stacking approach, as shown in Table 1. The algorithms' performance was evaluated using accuracy and the f1-score.

B. Performance Evaluation

We assessed the effectiveness of the improved prediction system's model using a set of criteria. Scikit-learn accuracy and f1-score were used to assess the performance of the machine learning algorithms. This is shown in Table 1. And Fig. 2 respectively. Logistic Regression had an accuracy score of 85.26%, K-Nearest Neighbor (KNN) had an accuracy score of 72.44%, Support ector Classifier (SVC), and the stacking method of the ensemble technique all had accuracy scores of 84.01% up to 85.52%. The algorithms had f1-scores of 85.26%, 75.64%, 84.89%, and 85.52%, respectively.

Table 1: Machine Learning Algorithm Result

Model	Accuracy	F1
Logistic Regression	85.26%	85.26%
KNN	72.44%	75.64%
SVC	84.01%	84.89%
Stacking Method	85.52%	85.52%

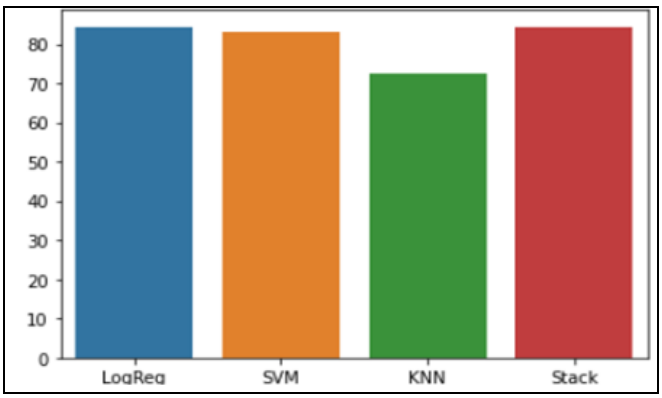

Fig 2: Accuracy of the Machine Learning Algorithms.

C. Physical Implementation/Deployment

The physical implementation of the Cyberbullying Detection System is divided into two major parts: the user interface or frontend developed with HyperText Markup Language (HTML), Cascading Style Sheets (CSS), Bootstrap, and JavaScript; the backend developed with the Flask framework, a lightweight Python web framework; and the machine learning prediction model built with Python (Pandas, Numpy, Sci-kit Learn, NLTK, Seaborn, and Matplotlib). It manages the logic and interactions between the user interface and the machine learning model. The system is deployed as a web application, making it accessible via web browsers. This allows for easy integration into social media platforms, providing real-time detection of cyberbullying. This deployment framework ensures that the cyberbullying detection system is robust and user-friendly, making it a valuable tool for curbing cyberbullying on social media platforms. Fig. 3 and Fig. 4 depicts aspects of the user interface.

Fig 3: Home Section of the User Interface

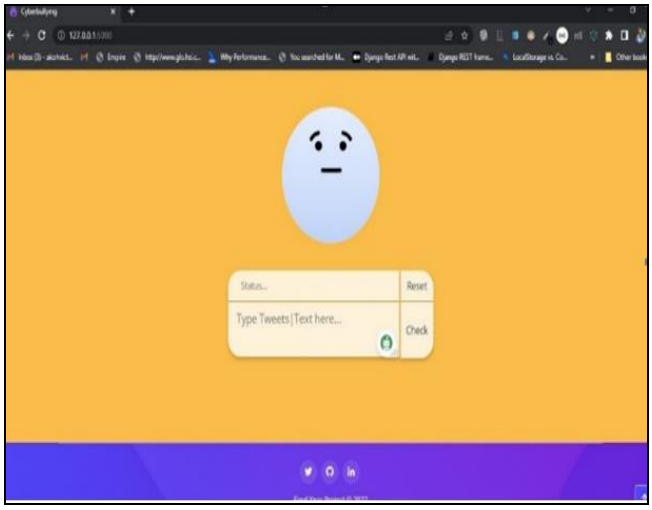

Fig 4: Interface for Post Input

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1058

V. CONCLUSION AND RECOMMENDATIONS

This research successfully developed an improved Cyberbullying Detection System by leveraging a stacking ensemble technique that combines Support Vector Machine (SVM), Logistic Regression (LR), and K-Nearest Neighbor (KNN) algorithms. The model achieved an accuracy of 85.52% on an X post dataset, demonstrating its effectiveness in detecting cyberbullying. The user interface, designed with HTML, CSS, Bootstrap, and JavaScript, coupled with a Flask-based backend, provides a user-friendly and accessible platform for real-time cyberbullying detection. This system can be integrated into social media applications to help mitigate the rising incidence of cyberbullying, making social media environments safer for users. Future research should focus on enhancing the model to detect cyberbullying in various languages such as pidgin English that are region-specific. Research can delve into addressing data imbalance issues to improve model robustness and exploring the effectiveness of the model across various social media platforms. Other research areas may include developing real-time deployment strategies to ensure seamless integration and operation as well as utilizing comprehensive evaluation metrics, including precision, recall, and user feedback, to provide a more dependable assessment of model performance.

REFERENCES

[1]. P. Ziman, C. Gaikwad, and A. Mhatre, (2021). “Detection of cyberbullying incidents on Instagram social network,” Intl. J. of Res. in Eng and Sci., vol. 9, pp. 6–13, 2021.

[2]. J. Mani, and J. P. Sainudeen, “A machine learning approach towards social media to tackle cyberbullying,” Intl. J. of Adv. Res. Id. and Inn. in Tech., vol. 4, pp. 495–498, 2018.

[3]. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, “Cyberbullying detection: hybrid models based on machine learning and natural language processing techniques,” Elctrncs, vol. 10, November 2021. https://doi.org/10.3390/electronics10222810

[4]. M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. AbdelMajeed, and T. Zia, “Abusive language detection from social media comments using conventional machine learning and deep learning approaches,” Mult. Sys., vol. 28, pp. 1925–1940, April 2021. https://doi.org/10.1007/s00530-021-00784-8

[5]. S. S. Jikriya, “Cyber bullying detection in social media using supervised ML & NLP techniques,” Intl. J. for Res. in App. Sc. and Eng. Tech., vol. 9, pp. 2259–2264, June 2021. https://doi.org/10.22214/ijraset.2021.35483

[6]. S. Kangane, P. Thorat, S. Indalkar, P. Yewale, and D. Deotale, “Detection of cyberbullying on social media using machine learning,” Intl. J. for Res. in Appd Sc. and Eng. Tech., vol. 9, pp.1401-1409, June 2022. https://doi.org/10.22214/ijraset.2021.38635.

[7]. Talpur, and D. O’Sullivan, “Cyberbullying severity detection: A machine learning approach,” PLOS ONE, vol.15, October 2020. https://doi.org/10.1371/journal.pone.0240924

[8]. J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, “Social media cyberbullying detection using machine learning. international journal of advanced computer science and applications,” Int. J. of Adv. Comp. Sc. and Appl., vol. 10, 2019. https://doi.org/10.14569/ijacsa.2019.0100587

[9]. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, “Automatic detection of cyberbullying in social media text,” PLOS ONE, vol. 13, October 2018. https://doi.org/10.1371/journal.pone.0203794

[10]. A. Kumar, “KNN Algorithm: When? Why? How? - towards data science,” Medium. https://towardsdatascience.com/knn-algorithm-what-when-why-how-41405c16c36f