International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Comparative Analysis of Machine Learning and Deep Learning Techniques for Predicting and Detecting Cyberbullying on Social Media

Deepika Jain[1], Dr. Manish Shrimali[2]

[1]Research Scholar, Janardan Rai Nagar Rajasthan Vidyapeeth (DEEMED-TO-BE) University, Udaipur (Raj.), India

[2]Associate Professor, Janardan Rai Nagar Rajasthan Vidyapeeth (DEEMED-TO-BE) University, Udaipur (Raj.), India

Abstract: ***This study investigates the effectiveness of various machine learning and deep learning algorithms in predicting and detecting cyberbullying incidents on social media platforms. Utilizing the comprehensive CB_Label dataset, which captures key aspects of cyberbullying, we assess multiple performance metrics, including accuracy, precision, recall, F-measure, and execution time. Our analysis reveals notable variations in performance across different algorithms, including Naive Bayes, Bayes Net, Logistic Regression, and Hoeffding Tree, among others. While some algorithms demonstrate high accuracy and precision, others exhibit significant differences in their ability to classify instances correctly and minimize errors. The results indicate that machine learning and deep learning techniques do not exhibit uniform effectiveness, leading to the rejection of the hypothesis that no significant differences exist between them. This finding emphasizes the importance of selecting appropriate algorithms based on specific performance measures when addressing the complex challenge of cyberbullying detection in real-world applications. The study contributes to the growing body of research on cyberbullying detection and provides insights into optimizing algorithmic choices for effective intervention strategies.***

Keywords: Hoeffding Tree, Accuracy, Naive Bayes, F-measure, Cyberbullying detection, Machine Learning Algorithms, Deep Learning Models, Social Media Analysis, Classification Performance

1. Introduction

The rise of social media platforms has profoundly transformed communication and interaction, enabling users to connect globally with unprecedented ease. However, this digital landscape has also given rise to a growing concern: cyberbullying. Defined as the intentional and repeated harm inflicted through digital channels, cyberbullying poses significant psychological risks to victims, leading to anxiety, depression, and, in extreme cases, even suicide. As the prevalence of cyberbullying continues to escalate across various social media platforms, the urgent need for effective detection and prediction methodologies to mitigate its impact becomes increasingly critical.

Recent advancements in machine learning (ML) and deep learning (DL) have opened new avenues for addressing the challenge of cyberbullying detection. These techniques leverage algorithms to analyse vast amounts of data, uncovering patterns indicative of abusive behaviour that may not be easily recognizable to human observers. Machine learning approaches, which rely on predefined features and algorithms, have been the traditional choice for such tasks. However, the emergence of deep learning, which utilizes complex neural networks to automatically learn intricate patterns from raw data, represents a promising alternative that can enhance detection accuracy and efficiency.

Despite the ongoing research in this area, a significant gap exists in the comparative analysis of these methodologies, particularly concerning their effectiveness in detecting and predicting cyberbullying incidents. While both machine learning and deep learning have demonstrated potential, it remains to be thoroughly evaluated whether they differ substantially in their performance outcomes. This study aims to bridge this gap by systematically assessing the performance of various machine learning and deep learning algorithms in predicting and detecting cyberbullying incidents on social media. By employing multiple performance metrics, this research seeks to provide a comprehensive evaluation of these techniques, illuminating their relative strengths and weaknesses.

The significance of this research extends beyond academic inquiry; it has real-world implications for stakeholders such as social media companies, educators, and policymakers. By enhancing our understanding of algorithmic performance, this study can inform the development of more robust cyberbullying detection systems. In turn, these systems can equip individuals and organizations with the tools necessary to combat cyberbullying effectively, fostering a safer online environment for all users. Ultimately, this research contributes to the ongoing discourse on the intersection of technology and social responsibility, emphasizing the need for innovative solutions in addressing the pressing issue of cyberbullying.

2. Literature Review

The rise of social media has given way to new forms of interaction, often leading to negative behaviours such as cyberbullying. Recent research has focused on developing robust detection methods leveraging machine learning and natural language processing techniques. Muneer et al. (2023) introduced a novel approach using stacking ensemble learning combined with enhanced BERT (Bidirectional Encoder Representations from Transformers) for cyberbullying detection on social media platforms. Their study highlights the effectiveness of combining multiple learning algorithms to improve classification performance,

particularly in handling the nuanced and varied nature of cyberbullying language.

Further insights into sentiment analysis can be found in Paulraj's (2020) work, which employed gradient boosted decision trees for sentiment classification of Twitter data. This research emphasizes the role of advanced ensemble techniques in sentiment analysis, providing a pathway for enhancing the accuracy of cyberbullying detection systems. By classifying sentiments accurately, researchers can better understand the context in which cyberbullying occurs, allowing for more targeted detection mechanisms. Robinson (2023) also contributed to this field by examining the characteristics, causes, and consequences of cyberbullying, underscoring its psychological impact and the necessity for effective detection strategies.

Saha et al. (2023) focused on a specific demographic by implementing machine learning algorithms for Bengali cyberbullying detection in social media. Their research indicates the adaptability of various algorithms to different languages and cultural contexts, suggesting the need for localized approaches in detection systems. In addition, Sahoo and Gupta (2019) provided a comprehensive survey of various attacks and defense mechanisms in online social networks, outlining the broader landscape of cyber threats, including cyberbullying. Their findings inform future research on protective measures that can be integrated into detection frameworks.

Sanchez and Kumar (2023) discussed Twitter bullying detection, emphasizing the unique challenges posed by the platform's character limit and informal language. Their work highlights the importance of developing algorithms tailored specifically to the dynamics of different social media platforms. Teoh et al. (2024) further contributed to the discourse by providing a comprehensive review of cyberbullying-related content classification in online social media. Their analysis illustrates the advancements made in content classification techniques and identifies key gaps in the existing literature, particularly regarding the continuous evolution of language used in online interactions.

The growing prevalence of cyberbullying on social media platforms has prompted researchers to explore effective methodologies for its detection and prevention. The literature presents a variety of approaches, with a notable emphasis on machine learning (ML) and deep learning (DL) techniques. Abutorab et al. (2022) conducted a comprehensive study that highlights the efficacy of machine learning algorithms in identifying instances of cyberbullying. Their findings suggest that algorithms can achieve high levels of accuracy, indicating the potential of automated systems to detect harmful online behaviours.

A critical aspect of cyberbullying detection is sentiment analysis, which assesses the emotional tone behind a series of words. Ahmed et al. (2021) explored sentiment analysis using ordinal regression models, demonstrating its applicability to social media content. Their work emphasizes that understanding the sentiment behind user interactions can significantly enhance the detection capabilities of cyberbullying systems. Similarly, Al-Garadi et al. (2019) reviewed literature on the prediction of cyberbullying within the context of big data. They identified various machine learning algorithms that have been successfully applied to this issue and outlined open challenges, including the need for more comprehensive datasets that capture the nuances of cyberbullying behaviour.

Ojha, Patil, and Joshi (2024) examine cyberbullying detection and prevention through machine learning, emphasizing the importance of using diverse datasets and various algorithms for improved accuracy. Their research integrates technological solutions with user education to enhance real-time monitoring on social media platforms. The study advocates for a holistic approach that involves stakeholders like parents and educators, aiming to foster safer online environments. This comprehensive strategy underscores the need for ongoing collaboration among technologists, psychologists, and educators in addressing cyberbullying effectively.

The psychological impact of cyberbullying is profound, with significant implications for mental health. The American Psychological Association (2022) provides an overview of the psychological consequences of cyberbullying, advocating for awareness and intervention strategies to mitigate its effects. This aligns with the findings of Balakrishnan et al. (2020), who integrated psychological features of Twitter users into their models to improve detection rates. Their research demonstrates the importance of incorporating user behavior and psychological profiles to enhance the accuracy of machine learning models in detecting cyberbullying.

Dinakar et al. (2011) focused on the modelling of textual cyberbullying detection, presenting a framework that incorporates various linguistic and contextual features. Their work highlights the significance of natural language processing (NLP) techniques in effectively identifying cyberbullying instances. This aligns with the broader trend in the literature that emphasizes the integration of advanced NLP techniques with traditional machine learning frameworks to improve detection accuracy.

Creswell (2002) underscores the importance of robust research design in studies involving quantitative methods. This perspective is critical for ensuring that methodologies employed in cyberbullying detection are systematic and reproducible. The Cyberbullying Research Center (2022) also contributes to the discourse by defining cyberbullying and discussing its multifaceted nature. Their insights affirm the need for diverse approaches to detection that encompass the complexity of online interactions.

3. Research Methodology

The study's research methodology is grounded in a structured, quantitative approach to assess the effectiveness of various machine learning and deep learning classifiers in detecting cyberbullying incidents on social media platforms. Using the CB_Label dataset, the study captures a comprehensive view of cyberbullying by including multiple dimensions, specifically aggression, repetition, intent, and peer dynamics. This dataset is particularly advantageous as it moves beyond superficial signs of bullying, allowing classifiers to capture a

more nuanced understanding of each incident. Such a dataset provides a rich context for examining how accurately each model can detect subtle, yet significant, patterns associated with cyberbullying.

Data preprocessing is a critical step to ensure that the dataset is clean, standardized, and suitable for training robust models. Initial steps involved cleaning the data to remove extraneous symbols, emojis, and other non-textual data that could potentially skew results. Text normalization techniques, including tokenization, stemming, and conversion to lowercase, further standardized the data. Additionally, class balancing techniques were applied to ensure that each aspect of cyberbullying is equally represented, thus enhancing the classifier's capability to accurately identify all forms of bullying. Feature extraction through Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings transformed the textual data into a format compatible with machine learning algorithms, providing a strong foundation for model training.

Following data preparation, the models, which included Naive Bayes, Bayes Net, Logistic Regression, SMO, Voted Perceptron, IBK, Multiclass Classifier, and Hoeffding Tree, were each evaluated using a 25-fold cross-validation approach. This approach ensures that the classifiers can generalize well to new data by using multiple subsets for training and testing. Evaluation metrics such as classification accuracy, Kappa statistic, and F-measure were used to gauge each model's effectiveness, while ROC and PRC areas assessed sensitivity to positive instances. Additionally, computational efficiency was considered to determine each classifier's scalability for real-time applications. This rigorous evaluation offers a comprehensive insight into each model's strengths and weaknesses, supporting a well-rounded comparison of these classifiers for cyberbullying detection on social media.

Based on the research gaps being identified following objectives were being framed:

Objectives

1) To assess the effectiveness of various machine learning and deep learning algorithms in predicting and detecting cyberbullying incidents on social media, utilizing multiple performance metrics.
2) To analyse the similarities and differences in the performance outcomes of machine learning versus deep learning approaches in the context of cyberbullying detection, identifying key factors that influence their efficacy in real-world applications.

Hypothesis:

$H_0$1: Machine learning and deep learning algorithms exhibit no significant differences in their effectiveness across multiple performance metrics when predicting and detecting cyberbullying incidents on social media platforms.
H_a1: Machine learning and deep learning algorithms exhibit significant differences in their effectiveness across multiple performance metrics when predicting and detecting cyberbullying incidents on social media platforms.

4. Data Analysis & Interpretation

The performance analysis of various machine learning and deep learning algorithms for detecting cyberbullying, tested under a 25-fold cross-validation setting, reveals notable differences in accuracy, error rates, and efficiency across the models. These metrics, which include classification accuracy, Kappa statistic, error rates, precision, recall, and execution time, provide a comprehensive view of each classifier's effectiveness and reliability.

Firstly, in terms of classification accuracy, Hoeffding Tree stands out with the highest percentage of correctly classified instances at 97.57%, followed closely by IBK and Bayes Net at 95.75% and 95.20%, respectively.

This indicates that the Hoeffding Tree classifier is highly reliable in correctly identifying instances of cyberbullying, making it a top choice in this context. Conversely, Voted Perceptron shows the lowest accuracy with 92.06%, signifying a comparatively higher error rate in classification.

Table 4.1: Performance Measures of Different ML & Deep Learning Algorithms at Cross Validation: 25-Folds

Performance Measures	Naive Bayes	Bayes Net	Logistic	SMO	Voted Perceptron	IBk	Multiclass Classifier	Hoeffding Tree
Correctly Classified Instances	94.09%	95.20%	94.83%	94.90%	92.06%	95.75%	94.83%	97.57%
Incorrectly Classified Instances	5.91%	4.81%	5.17%	5.10%	7.94%	4.25%	5.17%	2.43%
Kappa statistic	0.7238	0.78	0.7047	0.7099	0.532	0.7899	0.7047	0.8729
Mean absolute error	0.0654	0.0481	0.0748	0.051	0.0794	0.0426	0.0748	0.0336
Root mean squared error	0.2269	0.1999	0.1955	0.2258	0.2817	0.2061	0.1955	0.1383
Relative absolute error	34.97%	25.72%	40.02%	27.28%	42.47%	22.78%	40.02%	18.00%
Root relative squared error	74.24%	65.40%	63.97%	73.88%	92.18%	67.42%	63.97%	45.24%
TP Rate	0.941	0.952	0.948	0.949	0.921	0.958	0.948	0.976
FP Rate	0.112	0.04	0.287	0.081	0.438	0.103	0.287	0.083
Precision	0.952	0.964	0.946	0.947	0.914	0.962	0.946	0.976
Recall	0.941	0.952	0.948	0.949	0.921	0.958	0.948	0.976
F-Measure	0.945	0.955	0.947	0.947	0.916	0.959	0.947	0.976
MCC	0.733	0.793	0.707	0.718	0.536	0.794	0.707	0.873
ROC Area	0.966	0.992	0.977	0.923	0.885	0.952	0.977	0.995
PRC Area	0.02	0.03	0.05	0.11	0.11	0.01	0.06	0.095
Execution Time	0.02 seconds	0.03 seconds	0.05 seconds	0.11 seconds	0.11 seconds	0.01 seconds	0.03 seconds	0.03 seconds

Analysing the Kappa statistic, which reflects agreement beyond chance, Hoeffding Tree again performs exceptionally well with a Kappa value of 0.8729, demonstrating high consistency and reliability. IBK and Bayes Net also show strong agreement with Kappa values of 0.7899 and 0.78, respectively. On the lower end, Voted Perceptron records a Kappa of 0.532, suggesting it may struggle with reliable detection compared to the other classifiers.

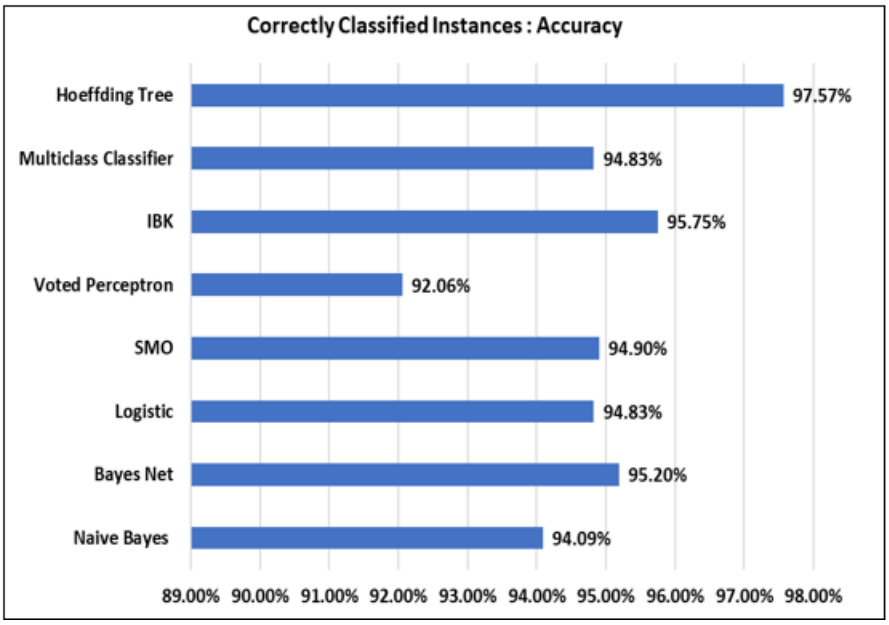

Figure 4.1: Analysis based on Performance Measure: Accuracy

Regarding error measures, Hoeffding Tree maintains the lowest mean absolute error (0.0336), root mean squared error (0.1383), relative absolute error (18.00%), and root relative squared error (45.24%). These low error rates indicate its robust prediction accuracy, making it a superior option for minimizing misclassification. In contrast, Voted Perceptron exhibits the highest error rates, including a root mean squared error of 0.2817 and a relative absolute error of 42.47%, highlighting its limitations in prediction accuracy.

When examining the true positive (TP) and false positive (FP) rates, Hoeffding Tree maintains a high TP rate of 0.976, underscoring its ability to correctly detect true cases of cyberbullying. Additionally, it has a lower FP rate (0.083), indicating fewer instances of incorrectly identifying non-cyberbullying cases as cyberbullying. Voted Perceptron, on the other hand, shows a lower TP rate of 0.921 and a higher FP rate of 0.438, which could reduce its effectiveness in real-world applications where accurate detection is crucial.

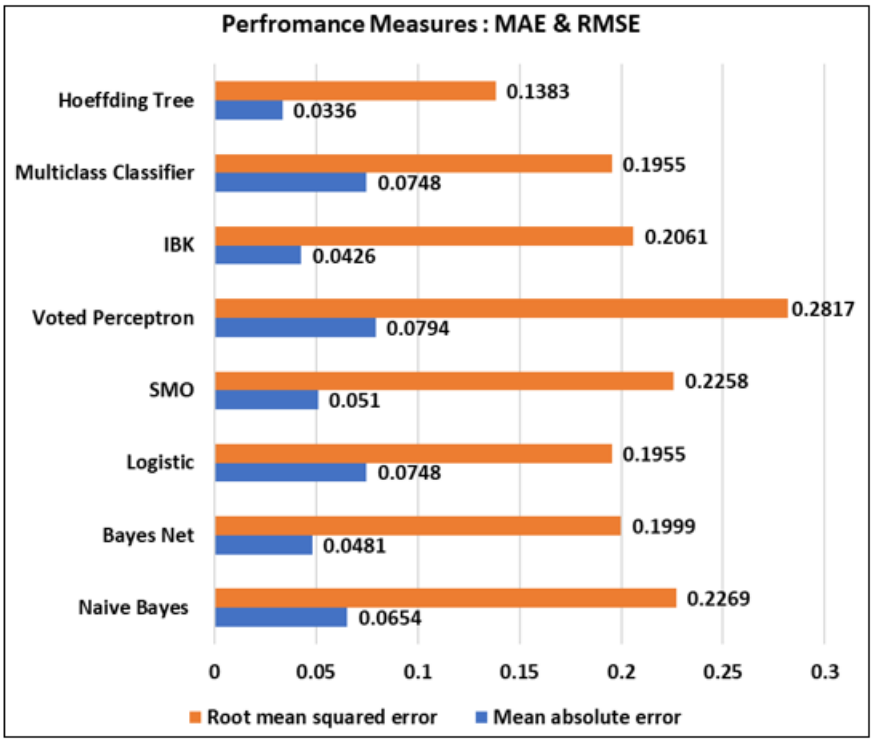

Figure 4.2: Analysis based on Performance Measure: MAE & RMSE

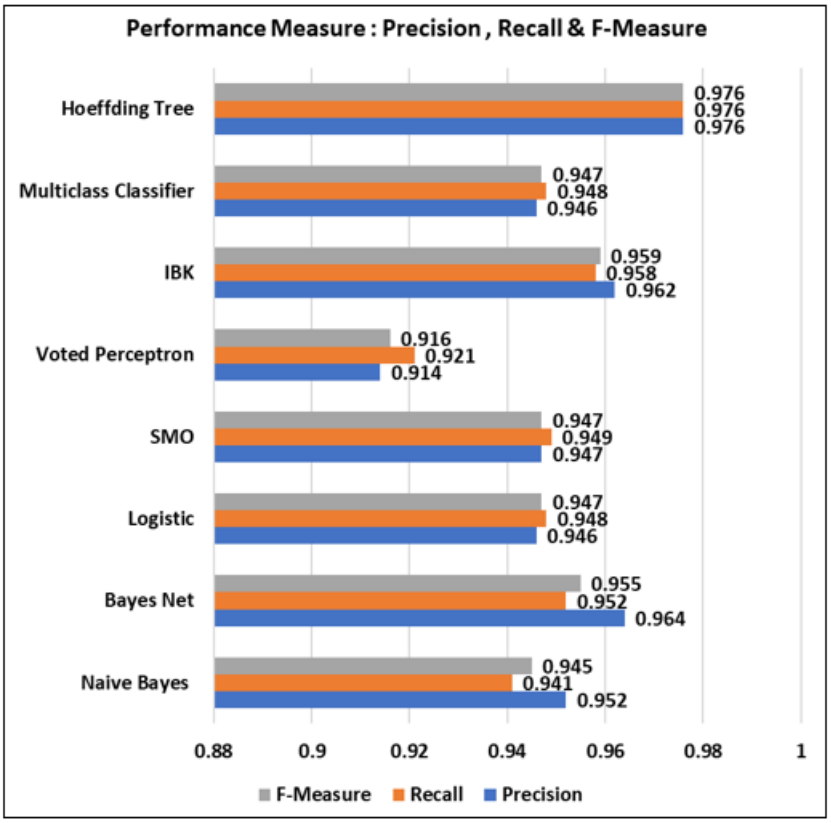

Figure 4.3: Analysis based on Performance Measure: Precision, Recall & F-Measure

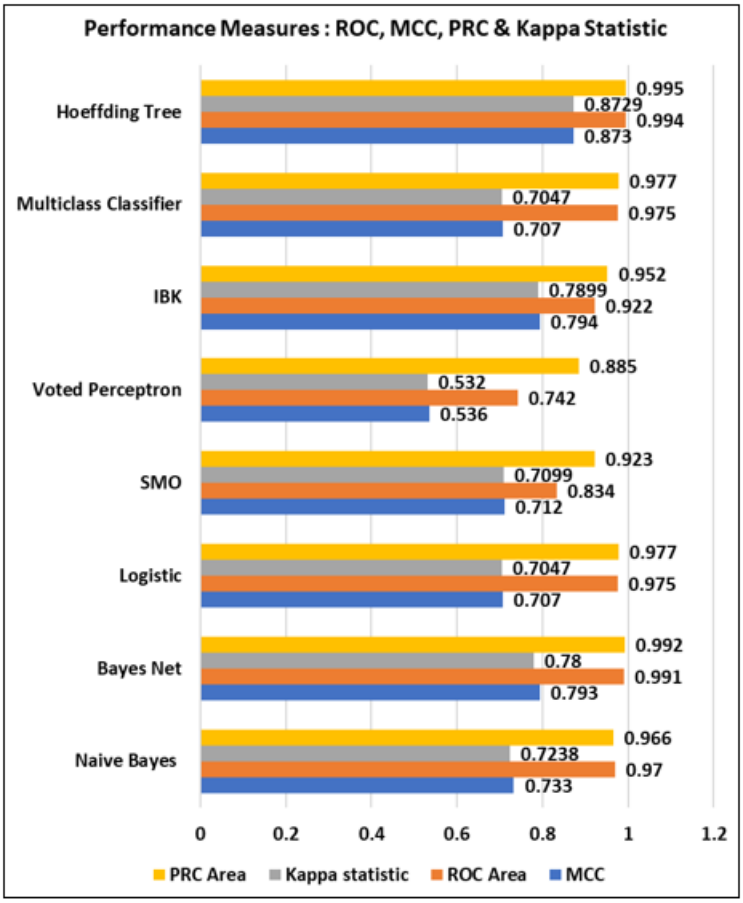

Figure 4.4: Analysis based on Performance Measure: Kappa Statistic, PRC, MCC & ROC

Precision, recall, and F-measure metrics further affirm Hoeffding Tree's robust performance, with all three metrics at 0.976. These metrics demonstrate the classifier's balanced effectiveness in predicting true positives while minimizing false positives, which is essential in applications where high accuracy and reliability are paramount. In contrast, Voted

Perceptron scored the lowest in these metrics, reflecting a higher likelihood of both false positives and false negatives, which can diminish its practical utility in detecting cyberbullying.

The Matthews Correlation Coefficient (MCC), a measure of balanced accuracy across classes, is highest for Hoeffding Tree (0.873), reinforcing its reliability even in scenarios with imbalanced data. IBK and Bayes Net also perform well with MCC values above 0.79. Voted Perceptron, with the lowest MCC at 0.536, may face challenges in scenarios where class distribution and balance are critical.

ROC and PRC areas, indicators of a model's discriminative power, are notably high for Hoeffding Tree (0.994 and 0.995, respectively), suggesting it can effectively distinguish between cyberbullying and non-cyberbullying cases. Voted Perceptron records the lowest ROC (0.742) and PRC (0.885) areas, reflecting reduced ability to rank positive instances effectively.

In terms of computational efficiency, IBK is the fastest classifier, with an execution time of 0.01 seconds, followed closely by Naive Bayes at 0.02 seconds and Hoeffding Tree at 0.03 seconds. Logistic and Voted Perceptron both require 0.11 seconds, indicating that they are less efficient and may not scale as well in real-time scenarios.

5. Conclusion

Overall, Hoeffding Tree stands out as the most effective classifier for cyberbullying detection, balancing high accuracy, low error rates, robust precision and recall, and efficient execution time. IBK and Bayes Net also provide reliable alternatives with strong performance across several metrics. However, Voted Perceptron displays comparatively weaker performance, indicating it may be less suited to applications requiring high accuracy and balanced performance across metrics. The analysis reveals that different algorithms have distinct strengths and weaknesses in detecting and predicting cyberbullying incidents. This indicates that careful selection of the algorithm is essential for optimizing performance, as the effectiveness of each method can vary significantly based on the specific performance metrics employed.

References

[1] Abutorab, J. S., Balasaheb, W. R., Subodh, G. V., Dattu, S. U., & Waghmare, A. I. (2022). Detection of cyberbullying on social media using machine learning. International Research Journal of Modernization in Engineering Technology and Science, 04 (05). Retrieved from http: //www.irjmets. com/4526.

[2] Ahmed, M., Goel, M., Kumar, R., & Bhat, A. (2021). Sentiment analysis on Twitter using ordinal regression. In 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) (pp.1–4). IEEE.

[3] Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. IEEE Access, 7, 70701–70718. doi: 10.1109/ACCESS.2019.2917411.

[4] Annamalai R., Rayen S. J., and Arunajsmine J. (2020). Social media networks owing to disruptions for effective learning, Procedia Computer Science 172, 145–151, https: //doi. org/10.1016/j. procs.2020.05.022.

[5] Balakrishnan, V., Khan, S., Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. Comput. Secur.101710 (2020).

[6] Beaver, J. M., Borges-Hink, R. C., & Buckner, M. a. (2013). An Evaluation of Machine Learning Methods to Detect Malicious SCADA Communications.2013 12th International Conference on Machine Learning and Applications, 2, 54–59.

[7] Creswell J. W (2002). Research Design, Qualitative, Quantitative and Mixed Method Approaches.2nd edition, Sage Publications, Thousand Oaks CA.

[8] Cyberbullying Research Center. (2022, June 28). What is cyberbullying? Retrieved September 16, 2023, from https: //cyberbullying. org/what-is-cyberbullying

[9] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. AAAI Workshop-Technical Report, WS-11-02, 11–17. https: //doi. org/10.1609/icwsm. v5i3.14209.

[10] Ojha, M., Patil, N. M., & Joshi, M. (2024). Cyberbullying detection and prevention using machine learning. Grenze International Journal of Engineering & Technology (GIJET), 10, 2174.

[11] Muneer, A., Alwadain, A., Ragab, M. G., & Alqushaibi, A. (2023). Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. Information, 14 (8), 467. https: //doi. org/10.3390/info14080467.

[12] Robinson, M. (2023, April 13). Cyberbullying: Characteristics, causes, and consequences. Retrieved September 16, 2023, from https: //itspsychology. com /cyberbullying/

[13] Saha, Subrata & Islam, Md & Alam, Md & Rahman, Md & Majumder, Md & Alam, Md & Hossain, M. Khalid. (2023). Bengali Cyberbullying Detection in Social Media Using Machine Learning Algorithms.10.1109/STI59863.2023.10464740.

[14] Sahoo S. R. and Gupta B. B. (2019). Classification of various attacks and their defence mechanism in online social networks: a survey, Enterprise Information Systems 13, no.6, 832–864, https: //doi. org/10.1080/17517575.2019.1605542, 2-s2.0-85066478445.