Leveraging High-Performance Computing for Enhanced Hate Speech and Offensive Language Detection on Twitter

Tahiatun Nazi
Dept of Computer Science and Engineering
tahiatun.nazi@g.bracu.ac.bd

Sheikh Yasir Hossain Katib
Dept of Computer Science and Engineering
sheikh.yasir.hossain.katib@g.bracu.ac.bd

Saiyeda Sabiha
Dept of Computer Science
saiyeda.sabiha@g.bracu.ac.bd

Mourika Nigar Mouny
Dept of Computer Science and Engineering
mourika.nigar.mouny@g.bracu.ac.bd

Annajiat Alim Rasel
Dept of Computer Science and Engineering
annajiat@gmail.com

Humaion Kabir Mehedi
Dept of Computer Science and Engineering
humaion.kabir.mehedi@g.bracu.ac.bd

Farah Binta Haque
Dept of Computer Science and Engineering
farah.binta.haque@g.bracu.ac.bd

I. Literature review

A dramatic change in computational power has been brought about by the combination of machine learning and High-Performance Computing (HPC), which goes beyond simple technological advancement to reshape the fundamental principles of research and practical applications. With a thorough examination of current research trends and emerging phenomena that highlight this field's enormous significance, this comprehensive literature review delves deeply into the complex aspects of this dynamic field [1].
The analysis about identifying hate speech in the dynamical social media environment, especially during major events like the COVID-19 pandemic, has received a lot of attention. By providing a carefully thought-out methodology that includes preprocessing, feature engineering, and a variety of machine learning classifiers, this study significantly advances this field.Strong preprocessing techniques are essential, as the paper emphasizes, because language in online platforms is context-dependent and varied. The use of ensemble techniques (Decision Trees, Stochastic Gradient Boosting) and classic classifiers (Logistic Regression, Support Vector Machine) shows a thorough investigation of approaches. It is easy to understand how effective these classifiers are by contrast: the two most notable particular theories are decision trees and stochastic gradient boosting. In addition to contributing to the scholarly discourse on hate speech detection, this work provides insightful practical information that can be used to improve online hate speech mitigation tactics[2].

With significant advancements in hardware capacities and the constant development of advanced algorithms, the convergence of machine learning and HPC has grown into an irreversible force. The unification of these elements has sparked a revolutionary movement that is advancing computational capabilities to unprecedented levels and transforming scientific inquiry and real-world problem-solving to an unprecedented degree[3].

The disruptive potential of machine learning within HPC has broken down the barriers between traditional disciplines, infiltrating a wide range of domains. Researchers have taken significant steps, using machine learning in financial forecasting, medical diagnosis, autonomous frameworks, scientific simulations, a wide range of other fields. This widespread infiltration has reshaped the bounds of opportunity, inspiring groundbreaking breakthroughs and creative approaches in a wide range of industries[4].

This article explores the role of advanced technologies such as artificial intelligence, high-performance computing, and machine learning in addressing the challenges posed by the COVID-19 pandemic. It highlights the applications of these technologies in predicting and containing the spread of the virus, leveraging person-specific data. The article discusses how high-performance computing and machine learning can be effectively used in analyzing COVID-19 data. It mentions the advantages and difficulties, including concerns about privacy and security.[5].

Using artificial neural networks, an ML surrogate provides accurate predictions at reduced time and costs. A web application on nanoHUB supports simulation techniques for homework and instruction. The tool has been found to enhance learning in materials science and engineering by providing a versatile and interactive simulation environment through the use of machine learning technology. Enhanced interactivity, real-time engagement, and anytime access help students visualize output variations with input changes and develop an understanding of system behavior.[6].

This article underscores the significance of ML in revolutionizing both the IT industry and academia, with a focus on transformations in applications, software, and hardware. It emphasizes the challenges in achieving Exascale performance in High-Performance Computing (HPC) also discusses a potential solution, proposing a new approach to system and software development to meet evolving computing power needs, including those for Artificial Intelligence applications. The article anticipates the integration of HPC systems with a readiness for emerging workloads[7].

This paper addresses the detection of offensive language on Twitter due to the increasing prevalence of hate speech. It focuses on challenges in automated detection within the distinctive language of social networks. The proposed solution utilizes Linear SVM and Naive Bayes algorithms to enhance existing experiments. Literature review emphasizes recent advances in hate speech detection, including offensive language identification and categorization. The study employs a Twitter dataset, normalizing data through preprocessing, and uses machine learning techniques like BERT, LSTM, and CNN. Results show Naive Bayes outperforms Linear SVM, achieving 92% accuracy and 95% recall. The study concludes that Naive Bayes is an effective alternative for detecting offensive language in tweets, surpassing Linear SVM and competing with complex models like BERT and CNN. The authors recommend further exploration of real-time tweet classification and extending these algorithms to classify various types of text in future research[8].

The paper addresses the surge of toxic online content, particularly hate speech, utilizing deep learning for hate speech identification on Twitter. The proposed solution involves an LSTM-based classification system distinguishing hate speech and offensive language, achieving an 86% accuracy. Employing word embeddings with LSTM and Bi-LSTM networks, the study emphasizes the importance of addressing hate speech for online safety. It proposes a sentiment analysis classification system, leveraging deep learning to overcome baseline model limitations, such as data imbalance, and enhance accuracy. Through experiments with various LSTM and Bi-LSTM models, including simple and stacked networks, the research achieves notable improvements, emphasizing the importance of detecting and removing toxic content for a safer online environment[9].

This paper addresses the critical task of offensive language detection in the low-resource Marathi language on social media, focusing on cyberbullying and offensive content on mental health. The study explores various pre-trained BERT transformer models for offensive speech detection, specifically evaluating their performance on the HASOC 2022 dataset. The models include MuRIL, MahaTweetBERT, MahaTweetBERT-Hateful, and MahaBERT. The paper introduces data augmentation from external hate speech datasets, such as HASOC 2021 and MahaHate, to enhance model performance. Key contributions include showcasing the superiority of Marathi Tweet BERT models, revealing the limitations of hateful BERT models, and demonstrating the effectiveness of combining multiple hate speech datasets for improved performance. This work contributes to offensive language detection in low-resource languages, particularly Marathi, and provides insights for future research in this domain[10].

The paper addresses automated offensive language detection on Twitter, focusing on challenges like unique language formats. It aims to enhance Linear SVM and Naive Bayes algorithm performance, compared with existing studies. The related works section highlights tasks from SemEval-2019 and explores deep learning techniques. The dataset comprises 24,783 tweets, emphasizing offensive and normal language. Naive Bayes outperforms Linear SVM in accuracy (92% vs. 90%) and recall (95% vs. 92%).The conclusion emphasizes the sensitivity of Linear SVM to data type and the challenges in parameter tuning, praising the simplicity and effectiveness of the Naive Bayes classifier. The study contributes to combating offensive language on social media, providing insights for future work, including exploring additional classification algorithms and real-time tweet classification[11].

This paper introduces a new way to classify different stages of bone cancer by using modified architecture that includes inception modules and convolutional neural networks. Leveraging a publicly available dataset from the 'Stanford ML' group, the proposed model attains an impressive 92.68% accuracy on the testing dataset. The paper highlights the significance of addressing bone cancer, a musculoskeletal disorder impacting patients' quality of life, and underscores challenges associated with existing detection methods like costly and tedious MRI scans. The proposed model, featuring four convolutional layers, two inception modules, and two fully connected layers, is detailed, emphasizing the advantages of inception modules in capturing multi-scale features while minimizing parameters. The paper also discusses data preprocessing and augmentation techniques to enhance model performance. The proposed model outperforms other techniques such as logistic regression, support vector machines, and random forests, according to evaluation metrics like accuracy, precision, recall, and F1-score. It achieves 92.68% accuracy and an impressive F1-score of 0.92 on the testing dataset[12].

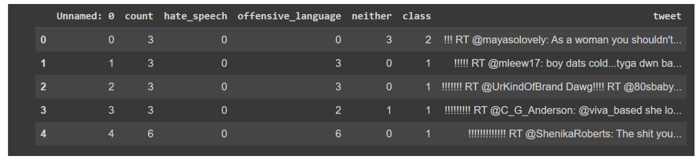
The paper investigates hate speech classification using machine learning methods and compares their efficacy with advanced deep learning models. Employing Twitter data labeled into hate speech, offensive, and neutral categories, the study explores logistic regression, random forest, naive Bayes, and support vector machine as machine learning methods, along with recurrent neural networks and bidirectional encoder representations (BERT) as deep learning models. Using accuracy, precision, recall, and F1-score as performance evaluation metrics, it is evident that both machine learning and deep learning models are effective in detecting hate speech. BERT achieves the highest overall accuracy at 87.78%, while support vector machines produce the best result (84.66%) among traditional classifiers. The paper discusses challenges and limitations of the methods and suggests future research directions[13].

The paper presents an automated classification approach for tweets on Twitter, categorizing them into hateful, offensive, or clean classes. The study evaluates different machine learning models like logistic regression, support vector machines, random forest, and k-nearest neighbors. It uses n-grams as features and measures their TFIDF values. Through a comprehensive analysis involving different n-gram values and TFIDF (frequency-inverse document frequency) normalization methods, the paper identifies the optimal model as a support vector machine with n=2 and l2 normalization, achieving an impressive 95.6 percent accuracy on test data, surpassing existing methods. Additionally, the paper introduces a module acting as an intermediary between users and Twitter, capable of filtering out hateful and offensive tweets from the user's timeline. The discussion encompasses challenges and limitations, such as the need for more data and handling sarcasm, along with ethical concerns regarding censorship and freedom of speech. The paper suggests future research directions, including the exploration of deep learning models, integration of sentiment analysis, and improvements to the user interface[14].

II. Collected Data

The dataset contains tweets labeled as hate speech, offensive language, or neither. The dataset consists of 24,802 tweets from Kaggle. Out of these, 3,972 are classified as hate speech, 20,298 as offensive language, and 432 as neither. The dataset is sourced from a research paper by Davidson et al. on automated hate speech detection. The dataset used to study hate speech detection was created using Twitter data. The text is categorized as neither, hate speech and offensive language. It's crucial to remember that this dataset contains content that may be interpreted as racist, sexist, homophobic, or just plain insulting given the nature of the study. "Fig. 7", even at the beginning of a sentence.

Figure Labels: Thus, this is a classification problem where the goal is to predict the class of a given tweet as one of the three categories mentioned above. 0,1,2 refer to hate speech, offensive, or neither in the dataset in the class column respectively. 10000 rows from the dataset are used and we used the class column and tweet column for our model.



	Unnamed: 0	count	hate_speech	offensive_language	neither	class	tweet
0	0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	4	6	0	6	0	1	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...

Fig. 1. Hate Speech and Offensive Language Dataset

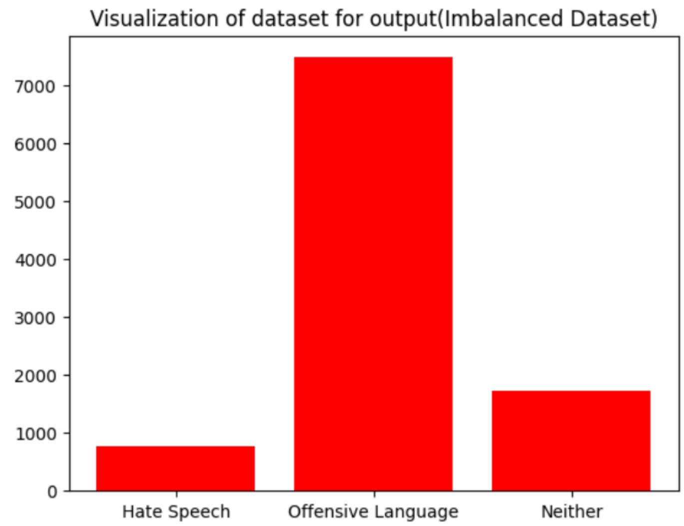


Fig. 2. Visualization of the Dataset

Figure Labels: The visualization of the frequency of data was done using the Matplotlib library to create a bar chart. This chart shows the frequency of each label category, which can give us an idea of how the tweets are distributed among the different categories.

III. Proposed Methodology

ML algorithms like Logistic Regression,Decision tree,Kth Nearest Neighbor,Naive Bayes Classifier Model to detect hate speech and offensive language are tested.These models were selected to capture various data patterns and improve overall prediction accuracy

The dataset used in this study is intended to help identify inflammatory language and hate speech in Twitter postings. It includes a wide range of elements, including user data, linguistic traits, and tweet content. Among the preprocessing processes were the handling of null values and the discretization of categorical characters. The dataset was split into training and test data using an 80% and 20% split.To forecast occurrences of hate speech and objectionable language, a number of machine learning methods were used, such as Decision Tree, Kth Nearest Neighbour (KNN), Naive Bayes Classifier, Logistic Regression, and Decision Tree. Furthermore, the model's evaluation made advantage of hidden values. We sought to forecast the categorization of a tweet with the following characteristics as a case study: User: @XYZUser, Language: English, Text: "This is an offensive tweet!" We applied the KNN technique to this problem, and our model was able to identify the offending text.Under the heading "Comparative Study of Machine Learning KNN, Naive Bayes Classifier, and Decision Tree Algorithm," the paper explores how to use these algorithms to identify hate speech and vulgar language. The simplicity of KNN comes from its ability to store the whole training dataset and predict new occurrences by comparing them to the K most similar ones. In binary classification, logistic regression performs exceptionally well, providing regularization, efficiency, and interpretability.Decision Trees are interpretable, manage non-linear relationships, highlight the significance of individual features, and serve as the foundation for effective ensemble techniques.Models are trained, and predictions are generated once the dataset is divided into training and testing sets. Effectiveness of the model is thoroughly assessed with suitable metrics designed for the identification of hate speech. As an example, essential measures are precision, recall, and F1-score. To improve performance, Decision Tree and KNN model hyperparameters are optimized. The final model for the detection of hate speech and objectionable language is determined by looking at which model shows the highest accuracy on the testing data.

K-Nearest Neighbor (KNN) KNN calculates the Euclidean distance (diagonal distance) between the query point and the number of the nearest neighboring points. then decides on the class label based on the label with the highest frequency. This classifier, which uses supervised learning, examines closeness to predict the classification or grouping of a particular data piece. Our software imports data from Scikit-Learn module KNeighborsClassifier from sklearn.neighbors, and model train with 3 being the default value of k. This produces an accuracy score of 93.0 percent after training on the dataset.

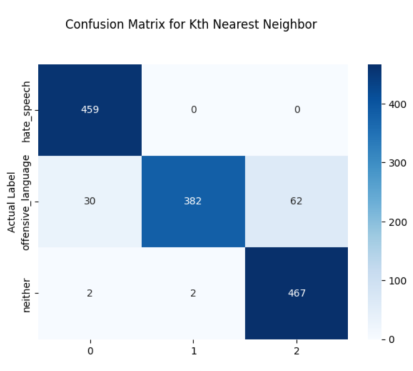


Fig. 3. Confusion Matrix of KNN

Logistic Regression A statistical model known as logistic regression, sometimes known as a supervised learning classifier, forecasts the likelihood that a binary event will occur on a given input dataset of independent variables. As a result, the dependent variable's output is a probability that ranges from 0 to 1 (inclusive). To train the model, we import LogisticRegression from sklearn.linear-model. When this is done, the accuracy on the dataset is 96.8 percent, which is better than the KNN model.

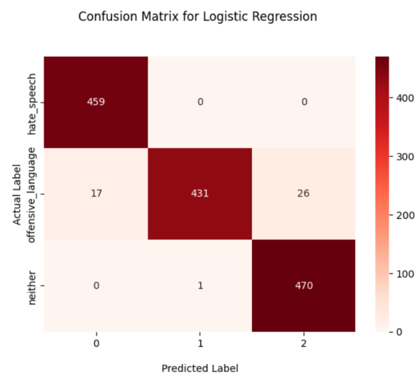


Fig. 4. Confusion Matrix of Logistic Regression

Decision Tree A decision tree, a supervised learning method, may be used to tackle classification and regression problems. It is a classifier with a tree-like structure, where each leaf node represents the classification result, each internal node a feature, and each branch decision point. To train the model using the decision tree learning approach, we import DecisionTreeClassifier from sklearn.tree. This yields a score of 96.3 percentage for accuracy, which is greater than that of the Naive Bayes model but lower than that of the Logistic Regression model.

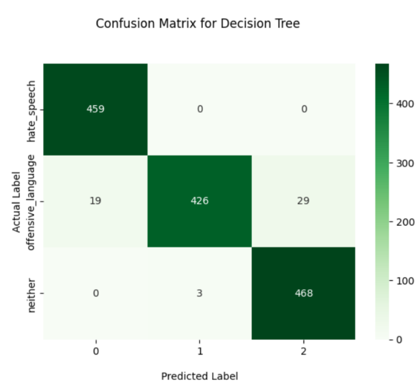


Fig. 5. Confusion Matrix of Desicion Tree

Naive Bayes Classifier in supervised learning statistics model known as the Gaussian Naive Bayes classifier was put into practice. It takes an input dataset of independent variables and uses it to estimate the probability of binary occurrences.

Trained the Naive Bayes model using the GaussianNB class from the sklearn.naive_bayes module.Utilized the dataset to train the Naive Bayes classifier, which produces a probability between 0 and 1 for the dependant variable.The classifier received an accuracy score of 91.6% when its accuracy on the dataset was assessed following model training.

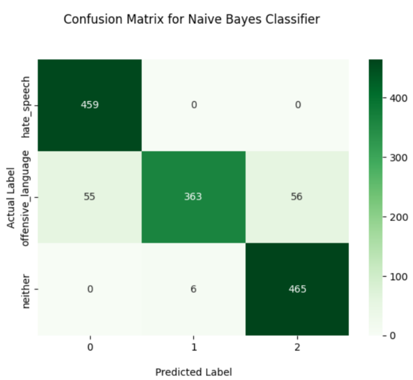


Fig. 6. Confusion Matrix of Naive Bayes Classifier

IV. Results And Analysis

These various models: KNN,Naive Bayes Classifier,Logistic regression and Decision tree were used for checking the accuracy rate of our analysis. The Logistic Regression model clearly has the highest accuracy score (96.8 percentage) after training on the dataset, while the Naive Bayes Classifier model has the lowest accuracy score (91.6 percentage), according to demonstrations of the four models. Last but not least, the Decision Tree model yields a score of 96.3 percentage, placing it in the middle of the KNN and Logistic Regression models. In light of the available dataset, the Logistic Regression model can detect hate speech and offensive language most accurately. Below, a bar chart that provides a clearer understanding of the outcomes serves as the visual depiction of the outcomes.

Model Name	Accuracy(%)
Logistic Regression	96.8
KNN	93.1
Decision Tree	96.3
Naive Bayes Classifier	91.6

Fig. 7.

References

[1] Y. Etsion and D. Tsafrir, "A short survey of commercial cluster batch schedulers," School of Computer Science and Engineering, Vol. 44221, The Hebrew University of Jerusalem, pp. 2005–2013, 2005.

[2] T. Khan, W. Tian, G. Zhou, S. Ilager, M. Gong, and R. Buyya, "Machine learning (ml)–centric resource management in cloud computing: A review and future directions," Journal of Network and Computer Applications, 2022, article 103405.

[3] T. Kurth, S. Treichler, J. Romero, M. Mudigonda, N. Luehr, E. Phillips, A. Mahesh, M. Matheson, J. Deslippe, M. Fatica, P. Prabhat, and M. Houston, "Exascale deep learning for climate analytics," in SC18: International Conference for High Performance Computing, Networking, Storage and Analysis, Nov 2018, pp. 649–660.

[4] Majeed, A., Lee, S. (2021). Applications of Machine Learning and High-Performance Computing in the era of COVID-19. Applied System Innovation, 4(3), 40.

[5] Jadhao, V., Kadupitiya, J. (2020). Integrating Machine Learning with HPC-driven Simulations for Enhanced Student Learning. 2020 IEEE/ACM Workshop on Education for High-Performance Computing (EduHPC), 25–34.

[6] Gepner, P. (2021). Machine Learning and High-Performance Computing Hybrid Systems, a New Way of Performance Acceleration in Engineering and Scientific Applications. Annals of Computer Science and Information Systems.

[7] Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach. ArXiv:1809.08651 [Cs]. https://arxiv.org/abs/1809.08651

[8] Hate Speech and Offensive Language Detection from Social Media — IEEE Conference Publication — IEEE Xplore. (n.d.). Ieeexplore.ieee.org. Retrieved December 8, 2023, from https://ieeexplore.ieee.org/document/9628255

[9] Shah, A., Singh, S. (2023). Hate Speech and Offensive Language Detection in Twitter Data Using Machine Learning Classifiers. 221–237. https://doi.org/10.1007/978-981-19-7455-7_17

[10] De Souza, G. A., Da Costa-Abreu, M. (2020). Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata. 2020 International Joint Conference on Neural Networks (IJCNN). https://doi.org/10.1109/ijcnn48605.2020.9207652

[11] Bisht, A., Singh, A., Bhadauria, H., Virmani, J., Kriti. (2020). Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model. ResearchGate, 243–264. https://doi.org/10.1007/978-981-15-2740-1_17

[12] Chavan, T., Patankar, S., Kane, A., Gokhale, O., Joshi, R. (n.d.). A Twitter BERT Approach for Offensive Language Detection in Marathi. https://arxiv.org/pdf/2212.10039.pdf

[13] De Souza, A., Abreu, D. (n.d.). Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata. https://shura.shu.ac.uk/26018/1/IEEE_WCCI_Artificial_Intelligence.pdf

[14] Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Malik, S. H. (2022). Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. International Journal of Information Management Data Insights, 2(2), 100120. https://doi.org/10.1016/j.jjimei.2022.100120