iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60148

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

AI Based Identification of Inappropriate Language

Padmaja Ragolu[1], Venkata Naveen Kumar Gollapalli[2], Sai Saranya Budumuru[3], Jayadeep Karagani[4], Vidhya Sai Vyshnavi Tripirineni[5]
Raghu Engineering College, India

Abstract: ***In the current digital era, the widespread use of online communication has raised the demand for automated systems that can recognize and block hate speech, improper language, and objectionable information. The usefulness of three different machine learning algorithms—Decision Trees, Random Forest, Long Short-Term Memory (LSTM), and others—in determining if a given text contains objectionable language is investigated in this work. While LSTM and represent state-of-the-art deep learning algorithms capable of processing unstructured text input, Random Forest and Decision Trees are standard machine learning techniques that rely on organized feature engineering. This study compares different algorithms in an effort to shed light on their advantages and disadvantages in dealing with the changing demands of online content moderation. Our results show each model's accuracy, precision, recall, and F1-score.***
Keywords: Random Forest, Decision Tree., LSTM .

I. INTRODUCTION

The omnipresent nature of online communication has given rise to a squeezing require for successful and effective strategies to recognize and combat the multiplication of improper dialect, despise discourse, and hostile substance on different online stages. As the volume of user-generated substance proceeds to take off, the manual control of content information has ended up an inconceivably challenge. Subsequently, the integration of machine learning calculations has developed as a promising arrangement to address this basic issue. This think about sets out on an investigation of the adequacy of four unmistakable machine learning calculations:
Choice Trees, Irregular Timberland, Long Short-Term Memory (LSTM), and , within the setting of foreseeing whether a given content contains hostile dialect. Choice Trees and Irregular Timberland, speaking to conventional machine learning strategies, depend on structured highlight designing to create forecasts. In differentiate, LSTM and stand at the bleeding edge of profound learning, prepared to handle unstructured content information, making them especially well-suited for the nuanced errand of hostile dialect location. The comparative investigation of these calculations points to shed light on their particular qualities and shortcomings, giving important experiences into their execution measurements such as exactness, exactness, review, and F1-score. By scrutinizing these measurements, this investigate points to evaluate the models' capacity to observe improper dialect viably. Eventually, the discoveries of this ponder hold the potential to catalyze the improvement of vigorous AI-based substance sifting frameworks. Such frameworks can essentially help online stages and social systems in keeping up secure and conscious online situations by naturally recognizing and hailing hostile content, in this way cultivating more comprehensive and capable advanced communities. By contributing to the creation of these frameworks, this inquire about tries to guarantee a more beneficial online talk for all clients, tending to the advancing challenges of online substance balance within the cutting edge computerized scene.

A. Objective of the Study

to evaluate and compare the effectiveness of three different machine learning algorithms, namely Decision Trees, Random Forest and Long Short-Term Memory (LSTM) in predicting the presence of offensive language in online text. Through the analysis of accuracy, precision, recall, and F1-score metrics, the study aims to provide insights into the strengths and weaknesses of these algorithms for content moderation purposes. Ultimately, the project seeks to contribute to the development of AI-based content filtering systems that can assist online platforms in creating safer and more respectful digital environments by automatically detecting and flagging offensive text.

B. Scope of the Study

This research focuses on evaluating the performance of three specific machine learning algorithms: Decision Trees, Random Forest and Long Short-Term Memory (LSTM) in identifying offensive language within text data. The study will assess their accuracy, precision, recall, and F1-score to gain insights into their capabilities and limitations. It does not delve into broader aspects of content moderation but concentrates on the technical aspect of text classification.

C. Problem Statement

In today's advanced scene, the fast development of online communication has driven to a surge within the dispersal of improper dialect, despise discourse, and hostile substance. This postures a noteworthy challenge for online stages in guaranteeing a secure and conscious online environment. The consider points to address this issue by assessing the viability of four particular machine learning calculations in distinguishing hostile dialect, giving important experiences for substance control.

II. RELATED WORK

Social media platforms like Twitter have become powerful tools for individuals to establish their reputation and promote ideas on a global scale. Twitter, distinct from other social media sources, is renowned for its capacity to disseminate real-time textual information. It serves as a vital platform for sharing opinions, ideas, and breaking news through succinct and compelling statements that reach millions worldwide. However, the immense reach of Twitter also poses potential risks, as posting inappropriate content can adversely affect the public image and privacy of individuals, including celebrities, politicians, and ordinary users. This paper aims to address the need for proactive measures to protect users' identities and reputations on Twitter by focusing on the automatic identification of potentially vulnerable messages. Our primary objective is to develop a classifier capable of predicting whether a user is likely to delete a particular post in the future. To achieve this, we employ Recurrent Neural Networks (RNNs), which leverage context-based information within tweets for effective classification. Additionally, our research contributes to the field by constructing an extensive set of features, including Twitter metadata, user information, and tweet text, to train classical machine learning algorithms on Twitter data. This work endeavors to enhance our understanding of post deletion behavior on Twitter and offers valuable insights into safeguarding users' online identities. [1]

To restore peace and harmony in this cross-cultural Internet era, it is of utmost importance for every citizen to behave and spread brotherhood. Under the given circumstances of 5G evolution citizens have taken their role onto the internet very seriously thereby most of the netizens spend their time condemning, judging, and trolling other netizens, public figures for that matter. Because of the consequences in an unprejudiced society involving race, gender, or religion, the challenge of automatically detecting hate speech and objectionable language in social media material is critical. However, existing research in this field is mostly focused on several languages, which limits its relevance to certain groups. The use of harsh language on social media platforms, as well as the consequences that this has, has become a serious problem in modern culture. Automatic ways to recognize and deal with this sort of content are necessary due to the large volume of content produced every day. Machine Learning & Natural Language processing has cutting-edge algorithms and classifiers that have benefitted mankind in impossible ways. Hence, our effort in this project is to make use of this impeccable technology to create an efficient system that automatically detects hate speech and offensive language from the Twitter dataset.[2]

Automated image classification has become a critical component in various medical and biological applications, enabling the efficient analysis of complex visual data. One such area of significance is the classification of HEp-2 (Human Epithelial type 2) cell patterns in immunofluorescence images, which plays a crucial role in diagnosing autoimmune diseases. The study presented in "Comparing convolutional neural networks and preprocessing techniques for HEp-2 cell classification in immunofluorescence images" by Rodrigues, Naldi, and Mari (2020) addresses this important domain. Immunofluorescence imaging offers valuable insights into cellular structures and can assist in the identification of distinct patterns associated with autoimmune diseases. The paper explores the application of Convolutional Neural Networks (CNNs), a powerful deep learning technique, and preprocessing techniques to enhance the accuracy of HEp-2 cell pattern classification. This research aims to contribute to the advancement of automated diagnostic tools, which can potentially revolutionize the efficiency and accuracy of autoimmune disease detection. In this introduction, we provide an overview of the significance of HEp-2 cell pattern classification, highlight the relevance of CNNs, and emphasize the study's objective of comparing these neural networks with preprocessing techniques. The paper's findings hold the promise of improving medical diagnosis and treatment through the integration of cutting-edge technology into the field of immunofluorescence image analysis.[3]

The paper titled "Accurate Disease Detection Quantification of Iris-Based Retinal Images Using Random Implication Image Classifier Technique" by Wang and Shan, published in Microprocessors and Microsystems in 2021, addresses a critical aspect of medical image analysis. The study focuses on enhancing disease detection and quantification in retinal images through an innovative approach known as the Random Implication Image Classifier Technique. Retinal diseases pose significant threats to vision and overall ocular health, making early and precise detection crucial for effective treatment and management. Traditional methods often encounter challenges in achieving high accuracy and reliability in disease detection from retinal images. To address this issue, the authors introduce a novel methodology based on the Random Implication Image Classifier Technique.

This paper not only presents an advanced technique but also provides insights into the potential applications and improvements in the field of medical image analysis. The research aims to contribute to the development of more accurate and efficient tools for the early diagnosis and quantification of retinal diseases, thereby potentially enhancing the quality of patient care and reducing the burden on healthcare systems.[4]

In the ever-evolving landscape of natural language processing and sentiment analysis, Mitra's 2020 paper titled "Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)" published in the Journal of Ubiquitous Computing and Communication Technologies (UCCT) delves into the application of machine learning techniques for sentiment analysis. Sentiment analysis, the process of discerning and quantifying emotional tone within text data, has gained prominence due to its wide-ranging applications, from market research to social media monitoring. Mitra's research focuses specifically on sentiment analysis using a lexicon-based approach, leveraging a dataset comprised of movie reviews. The choice of the dataset is significant as movie reviews often contain rich and nuanced expressions of sentiment, making them a valuable resource for studying sentiment analysis methodologies. The paper is set against the backdrop of the increasing relevance of sentiment analysis in the digital age, where understanding public opinion and sentiment on various topics is crucial. By harnessing machine learning techniques and a specialized lexicon, the study seeks to contribute to the advancement of sentiment analysis methodologies, shedding light on the practical applications of such techniques in the realm of movie reviews and potentially extending their utility to broader domains of text analysis. This paper's findings hold relevance for researchers, practitioners, and enthusiasts interested in the ever-expanding field of sentiment analysis and its real-world implications.[5]

III. PROPOSED SOLUTION

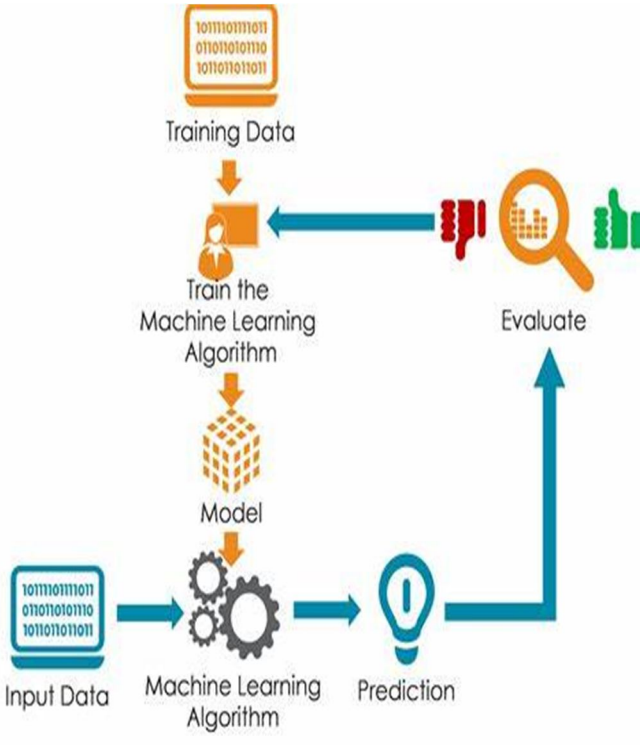


Figure 1 Proposed Block Diagram

This study employs advanced machine learning algorithms—Decision Trees, Random Forest, and —to identify and mitigate offensive language in online content, addressing a critical need for effective content moderation. The research evaluates these algorithms' capabilities in handling the intricacies of content moderation by analyzing their performance metrics: accuracy, precision, recall, and F1-score. Decision Trees offer a straightforward approach but may lack complexity for nuanced language, whereas Random Forest improves on Decision Trees by incorporating multiple decision trees to enhance the decision-making process, potentially offering better generalization. , a deep learning model based on transformer architecture, excels in understanding the context of words in sentences, which is crucial for detecting subtleties in offensive language.

By rigorously assessing these algorithms, the study aims to highlight their respective strengths and weaknesses in content moderation. The evaluation focuses on their effectiveness in distinguishing offensive content accurately, ensuring high precision to minimize false positives, and achieving significant recall to reduce false negatives, all quantified through the F1-score. This research contributes to the ongoing development of AI-based content filtering systems, aiming to create safer digital communities by providing insights into the efficacy of these machine learning solutions in combating offensive online content.

Enhanced Accuracy: By utilizing advanced machine learning models like , the system achieves high accuracy in identifying offensive language, reducing the prevalence of harmful content.

1) *Complex Understanding:* 's ability to understand the context of words in sentences allows for a more nuanced detection of offensive language, surpassing simpler keyword-based filters.
2) *Automated Moderation:* Automates the content moderation process, significantly reducing the manual effort required and enabling real-time content filtering.
3) *Scalability:* Machine learning models can handle vast amounts of online content efficiently, making the system highly scalable and suitable for platforms of any size.
4) *Continuous Learning:* The algorithms can be trained on new datasets to adapt to evolving language and slang, ensuring the system remains effective over time.
5) *Precision and Recall Balance:* Careful tuning of models ensures a balance between precision (minimizing false positives) and recall (minimizing false negatives), optimizing the moderation process.
6) *Versatility:* The system can be applied across various digital platforms, including social media, forums, and chat applications, to foster safer online communities.
7) *Reduced Bias:* Through rigorous training and evaluation, the models aim to minimize bias in content moderation, promoting fairness and equity in online interactions.
8) *User Experience Improvement:* By effectively filtering out offensive content, the system contributes to a more positive and respectful online environment, enhancing user satisfaction.

Insightful Analytics: Provides valuable insights into the nature and frequency of offensive content, aiding in the development of better content guidelines and policies.

IV. METHODOLOGIES

A. *Long Short-Term Memory (LSTM)*

Long Short-Term Memory (LSTM) is a recurrent neural network (RNN) architecture designed to address the vanishing gradient problem, a common issue in traditional RNNs. LSTM was introduced by Sepp Hochreiter and Jürgen Schmid Huber in 1997 and has since become a fundamental building block in various applications of deep learning, particularly in natural language processing, speech recognition, and time series analysis.

LSTMs are designed to capture and learn long-range dependencies in sequential data, making them well-suited for tasks involving sequences, such as text and speech processing, and time series prediction. They achieve this by introducing specialized memory cells that can store and update information over extended periods.

B. *Key components of an LSTM cell include*

Cell State (Ct): Represents the memory of the cell and can be updated, read, and written to selectively.

Hidden State (ht): Captures the output of the cell and can be thought of as the current "state of mind" of the LSTM.

Gates (Input, Forget, Output): These gates regulate the flow of information into and out of the cell, allowing LSTMs to control what information is retained or discarded over time.

The LSTM's ability to preserve information over long sequences and effectively handle vanishing gradients has made it a crucial tool in many deep learning applications, enabling better performance in tasks like machine translation, sentiment analysis, and speech recognition.

C. *Random Forest*

The Random Forest algorithm represents a significant advancement in the field of machine learning, particularly in solving regression and classification problems. At its core, Random Forest utilizes ensemble learning, a method that combines multiple classifiers to solve complex problems more effectively than any single classifier could alone.

This algorithm operates through the creation of numerous decision trees during training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Random Forests build on the concept of decision trees, which are simple yet powerful decision support tools that model decisions and their possible consequences as a tree-like structure. Decision trees comprise decision nodes, leaf nodes, and a root node, where the decision nodes represent the questions or tests on features, leaf nodes represent decisions or outcomes, and the root node is the feature that initiates the decision process.

The Random Forest algorithm enhances decision trees by creating a 'forest' through the method of bagging or bootstrap aggregating. This technique involves training each tree on a random subset of the data with replacement, ensuring that each tree learns from a different slice of the data, which reduces overfitting and increases model robustness. The algorithm then aggregates the predictions from all trees to decide on the final output. This aggregation improves prediction accuracy and helps mitigate the overfitting issue common in individual decision trees by averaging out biases.

One of the defining characteristics of Random Forest is its simplicity in handling missing values and its ability to balance errors in unbalanced datasets. Moreover, it requires minimal hyper-parameter tuning, making it an efficient model right out of the box, especially with libraries like Scikit-learn providing straightforward implementations.

D. The Random Forest algorithm offers several advantages

1) Accuracy: By combining the predictions from multiple decision trees, Random Forest typically achieves higher accuracy than individual decision trees, especially on complex datasets.
2) Handling Missing Data: It can maintain accuracy even when a significant portion of the data is missing.
3) Minimal Hyper-parameter Tuning: Unlike many other machine learning algorithms, Random Forest can produce very competitive results with its default settings.
4) Overfitting: Through the mechanism of bootstrapping and averaging predictions, it greatly reduces the risk of overfitting, which is a common problem with decision trees.
5) Variable Importance: Random Forest provides useful insights into which variables are important in the prediction process, offering a natural form of feature selection.
6) Versatility: Capable of performing both regression and classification tasks, it is also suitable for identifying outliers and clustering.
7) Handling Large Data Sets: The algorithm is exceptionally efficient, capable of handling thousands of input variables without variable deletion, and is faster than other ensemble techniques.
8) Flexibility: It can handle both numerical and categorical data without the need for scaling or normalization.
9) Ease of Use: Libraries like Scikit-learn have democratized the use of Random Forest, making it accessible to non-experts.
10) Parallelizable: The independent nature of trees in a Random Forest makes the algorithm highly parallelizable, leading to significant speedups in training.

In the decision-making process, entropy and information gain are pivotal. Entropy measures the disorder or unpredictability in the data, and information gain measures the reduction in entropy after the dataset is split on an attribute. Decision trees use these measures to decide where to split the data to achieve the most homogeneous sub-nodes.

In summary, the Random Forest algorithm is a powerful and versatile machine learning technique. Its ability to produce highly accurate predictions with minimal need for hyper-parameter tuning and its robustness to overfitting make it a popular choice for a wide range of problems. By leveraging the strengths of multiple decision trees and employing techniques like bagging, Random Forest addresses many of the limitations inherent in single decision trees, offering a more reliable and accurate prediction model that is highly valued across diverse applications.

a) Core Innovations of Bidirectional Context's

foremost innovation is its bidirectional processing capability. Traditional models processed text linearly, either from left to right or vice versa. This approach inherently restricted their ability to understand context fully. overcomes this by examining the context on both sides of a word simultaneously. This bidirectional context understanding significantly enhances the model's grasp of sentence structure and meaning, allowing for more nuanced interpretations of language.

b) Pre-training and Fine-tuning

operates in two stages: pre-training and fine-tuning. During pre-training, learns language patterns from a large text corpus without specific task-oriented training. It does this by predicting words that are randomly masked in sentences, gaining a deep understanding of language structure and context. In the fine-tuning stage, is adjusted for specific NLP tasks such as text classification, question answering, and named entity recognition. This approach allows to achieve state-of-the-art performance across various NLP challenges.

c) Transformer Architecture

At the heart of is the transformer architecture, which utilizes self-attention mechanisms to determine the relevance of each word in a sentence to every other word. Unlike previous architectures that processed text sequentially, transformers parallelize computation, allowing for more efficient learning of word relationships. This architecture underpins 's ability to understand long-range dependencies in text, a critical factor in its success.

d) Multilingual Capabilities

's design facilitates its adaptation to multiple languages, enabling its deployment in diverse linguistic settings. This multilingual capacity has broadened the reach of NLP applications, making advanced language models accessible to a global audience.

e) Impact of Enhanced Model Accuracy

By understanding the context more deeply, has significantly improved the accuracy of NLP models. Tasks such as sentiment analysis, machine translation, and content summarization have seen marked improvements in performance, leading to more reliable and nuanced AI-driven language applications.

f) Versatility Across NLP Tasks

's architecture allows for its application across a wide range of NLP tasks without substantial model restructuring. This versatility has made a foundational model in NLP, adaptable to various challenges from automated customer support to content curation.

g) Challenges and Future Directions

While has made substantial contributions to NLP, it is not without challenges. The model requires significant computational resources for training, making it less accessible for smaller organizations. Additionally, while 's context understanding is advanced, there remains room for improvement in areas like commonsense reasoning and ambiguity resolution.

Ongoing research aims to address these challenges, building on 's framework to develop even more sophisticated language models. Efforts to make models more efficient and to further enhance their understanding of nuanced language continue to push the boundaries of what's possible in NLP.

represents a major leap forward in natural language understanding, laying the groundwork for future advancements in AI. Its comprehensive approach to context, innovative training methodology, and transformative impact across a variety of NLP applications underscore its significance in the ongoing evolution of language models.

E. Decision Tree

The Decision Tree algorithm is a versatile and widely-used supervised learning method for classification and regression tasks, valued for its ability to represent data hierarchically. It operates by recursively dividing the dataset into purer subsets based on informative features, using criteria like Gini impurity or entropy for classification, and mean squared error for regression. This process creates a tree-like structure of nodes representing decisions or tests, and branches that denote the outcome of these decisions, culminating in leaf nodes that hold predictions.

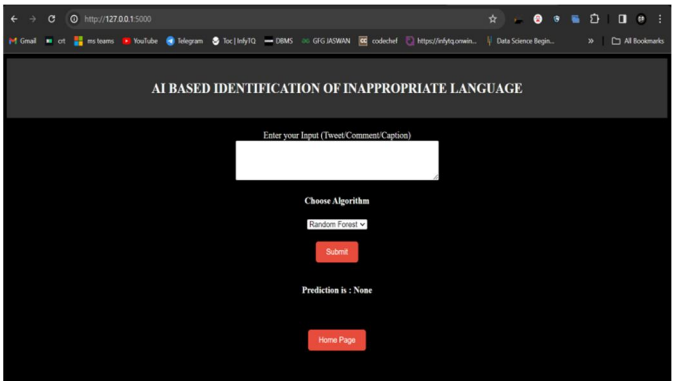
One of the main advantages of Decision Trees is their interpretability; the model's decisions can be easily traced as if-else rules from the root to the leaf nodes, making them particularly appealing in sectors requiring explainability, such as finance and healthcare. However, they are susceptible to overfitting, particularly with deep trees, which can be mitigated through pruning or setting maximum depth limits.

Decision Trees also serve as foundational elements for ensemble techniques like Random Forest and Gradient Boosting, enhancing predictive performance and robustness. Their ability to adeptly handle both numerical and categorical data underpins their broad application across various industries, embodying a balance of simplicity and effectiveness in machine learning.

V. RESULT AND DISCUSSION

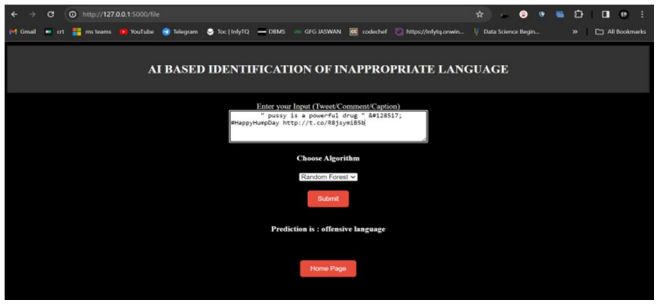
Home Page:

Here user view the home page of Inappropriate Language Prediction web application.



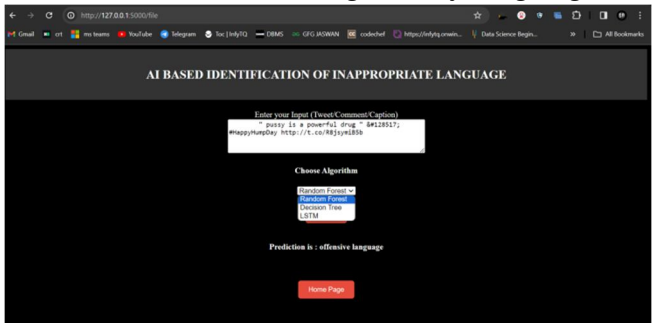
Page After Entering Input

Here we can see the webpage after entering the input.



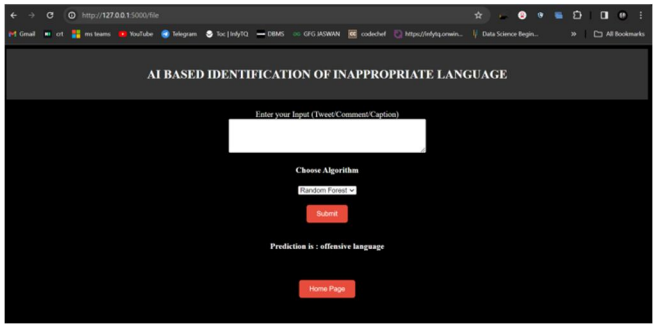
Algorithm Selection

Here user will click the select the algorithm by using drag down box.



Submit

After clicking on the submit button.The final prediction is displayed.

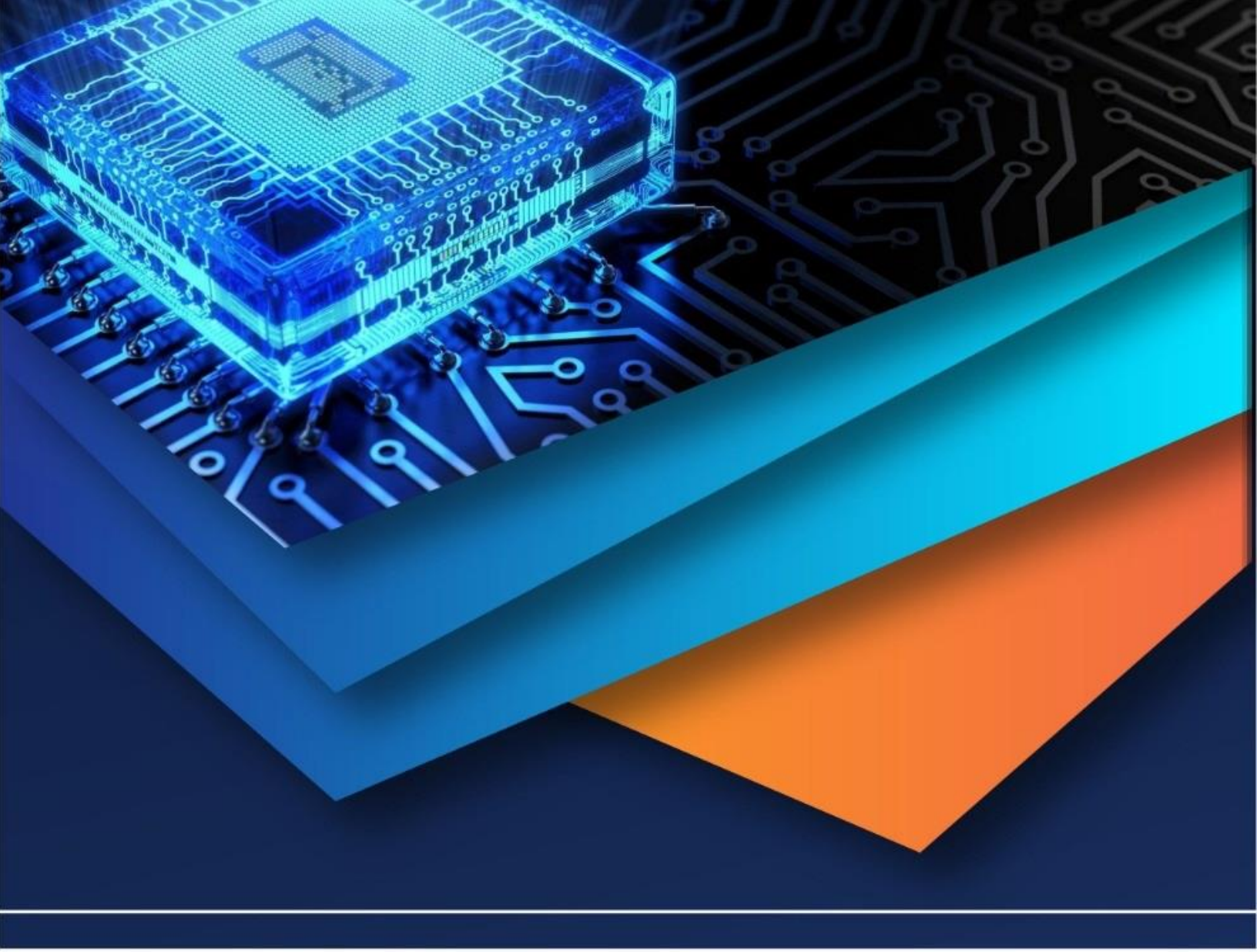


VI. CONCLUSION

In conclusion, this study has provided valuable insights into the effectiveness of four distinct machine learning algorithms – Decision Trees, Random Forest, LSTM, and – in detecting offensive language within online content. Each algorithm exhibited its unique strengths and weaknesses, with Decision Trees and Random Forest showcasing the importance of structured feature engineering, while LSTM demonstrated the power of deep learning techniques in handling unstructured text data. By evaluating accuracy, precision, recall, and F1-score, we have gained a comprehensive understanding of their predictive capabilities. The findings of this research hold significant promise for the development of robust AI-based content moderation systems, offering the potential to create more inclusive and respectful online spaces, thereby contributing to a healthier and more responsible digital discourse for all users.

REFERENCES

[1] Pandian, A. Pasumpon. "Performance Evaluation and Comparison using Deep Learning Techniques in Sentiment Analysis." Journal of Soft Computing Paradigm (JSCP) 3, no. 02 (2021): 123-134.
[2] Manoharan, J. Samuel. "Study of Variants of Extreme Learning Machine (ELM) Brands and its Performance Measure on Classification Algorithm." Journal of Soft Computing Paradigm (JSCP) 3, no. 02 (2021): 83-95.
[3] Ranganathan, G. "A Study to Find Facts Behind Preprocessing on Deep Learning Algorithms." Journal of Innovative Image Processing (JIIP) 3, no. 01 (2021): 66-74.
[4] Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, "Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach," 2019.
[5] Akhter, M. P., Jiangbin, Z., Naqvi, I. R., Abdelmajeed, M., Mehmood, A., & Sadiq, M. T. (2020). Document-level text classification using single-layer multisize filters convolutional neural network. IEEE Access, 8, 42689-42707.
[6] K. J. Madukwe and X. Gao, "The Thin Line Between Hate and Profanity," in Australasian Joint Conference on Artificial Intelligence, 2019, pp. 344–356.
[7] Beeravolu, A. R., Azam, S., Jonkman, M., Shanmugam, B., Kannoorpatti, K., & Anwar, A. (2021). Preprocessing of Breast Cancer Images to Create Datasets for Deep-CNN. IEEE Access, 9, 33438-33463.
[8] Chen, Z., Zhou, L. J., Da Li, X., Zhang, J. N., & Huo, W. J. (2020). The Lao text classification method based on KNN. Procedia Computer Science, 166, 523-528.
[9] Diker, A., Avci, E., Tanyildizi, E., & Gedikpinar, M. (2020). A novel ECG signal classification method using DEA-ELM. Medical hypotheses, 136, 109515.
[10] Heidari, M., Mirniaharikandehei, S., Khuzani, A. Z., Danala, G., Qiu, Y., & Zheng, B. (2020). Improving the performance of CNN to predict the likelihood of COVID19 using chest X-ray images with preprocessing algorithms. International journal of medical informatics, 144, 104284.
[11] Poloni, K. M., de Oliveira, I. A. D., Tam, R., Ferrari, R. J., & Alzheimer's Disease Neuroimaging Initiative. (2021). Brain MR image classification for Alzheimer's disease diagnosis using structural hippocampal asymmetrical. attributes from directional 3-D logGabor filter responses. Neurocomputing, 419, 126-135.
[12] Rodrigues, L. F., Naldi, M. C., & Mari, J. F. (2020). Comparing convolutional neural networks and preprocessing techniques for HEp-2 cell classification in immunofluorescence images. Computers in biology and medicine, 116, 103542



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY