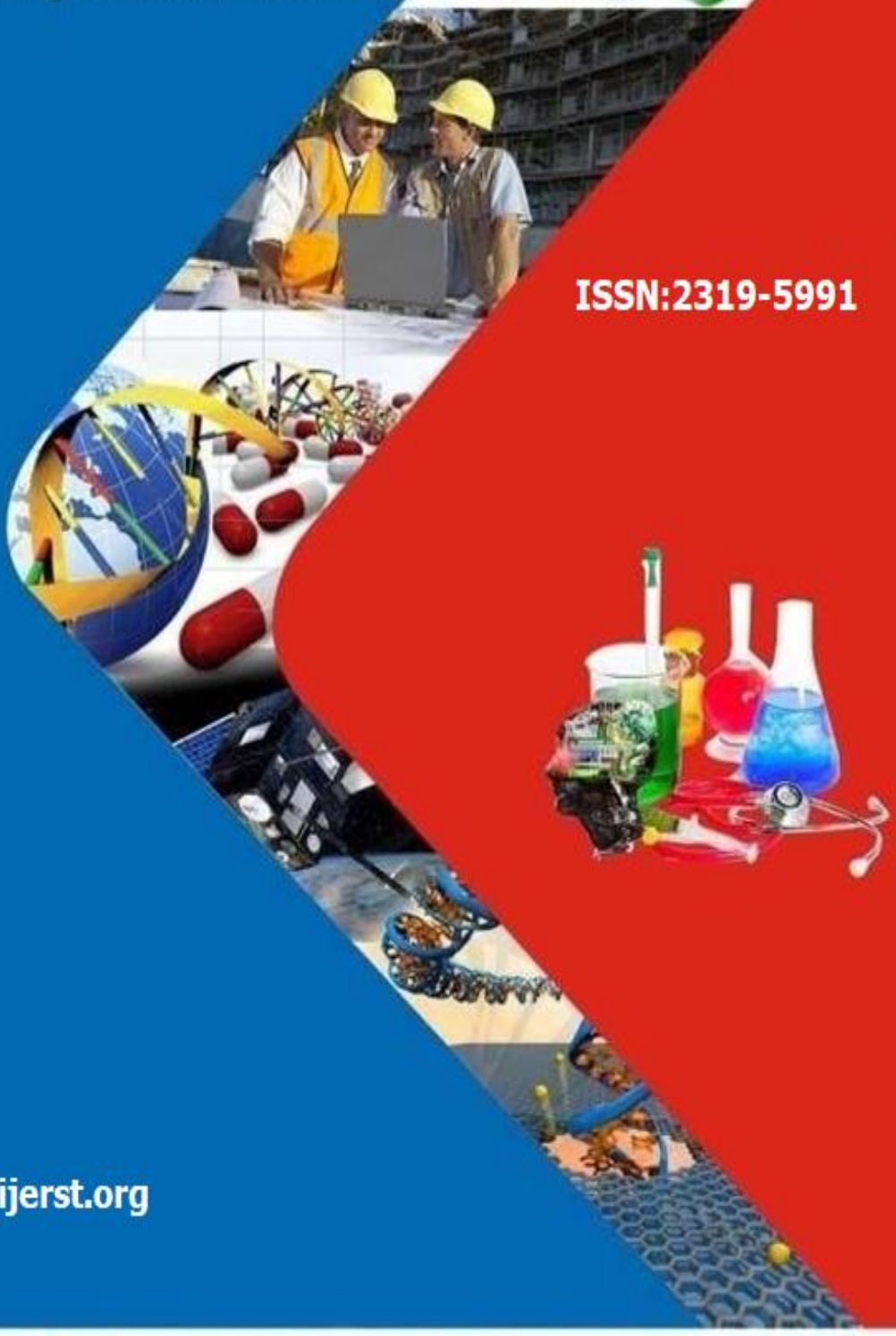
IJERST

International Journal of Engineering Research and Science & Technology

ISSN:2319-5991



www.ijerst.org

E-mail: editor@ijerst.org or ijerst.editor@gmail.com

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

MULTI-STAGE MACHINE LEARNING AND FUZZY APPROACH TO CYBER-HATE DETECTION

M. PRIYANKA[1], Y.GOPI SURESH[2]

[1] Assistant professor, Department of Master of Computer Applications, SRK Institute of Technology, Vijayawada, Andhra Pradesh

[2] MCA Student, Department ofMaster of Computer Applications, SRK Institute of Technology, Vijayawada, Andhra Pradesh

ABSTRACT

Multi-Stage Machine Learning and Fuzzy Approach to Cyber-Hate Detection aims to address the growing concern of hate speech and cyber bullying on social media platforms, where millions of users post content daily. Statistically, studies show that over 40% of internet users have experienced some form of online harassment, and nearly 70% of reported cases on social media go unaddressed or are delayed due to manual moderation. Traditionally, the manual system for detecting cyber hate involves human moderators reviewing flagged content based on user reports. This process is labor-intensive, time-consuming, subjective, and inconsistent due to the sheer volume of data and varying interpretations of hate speech across cultures and contexts. These limitations lead to delays in response and failure to filter out subtle or context-specific abusive content. Motivated by the need for faster, scalable, and more accurate solutions, this research proposes an intelligent system that combines a multi-stage machine learning pipeline with a fuzzy logic approach. The objective is to enhance detection accuracy and reduce ambiguity by capturing not only direct hate expressions but also context-based and indirect abusive content. In the proposed system, textual data undergoes preprocessing followed by multiple stages of classification using traditional machine learning algorithms like Naive Bayes, SVM, and Decision Tree, and then evaluated against a logistic regression model to determine the most reliable classifier. Furthermore, a fuzzy inference system is integrated to handle linguistic uncertainty and context sensitivity, allowing the system to make better decisions in edge cases where text may not be explicitly hateful but potentially harmful. This hybrid model leverages the strength of both deterministic and fuzzy learning methods to create a reliable and efficient cyber hate detection system, addressing the inefficiencies of manual approaches and contributing toward safer online communication environments.

Keywords: Machine Learning, Fuzzy Approach, Naïve Bayes, SVM.

1. INTRODUCTION

1.1 Background and Overview

Cyber-hate detection has become a crucial area of concern in India with the explosive growth of social media usage. As per a 2023 report by the National Crime Records Bureau (NCRB), cyberbullying and online harassment cases in India increased by 36% over the past two years. The accessibility of digital platforms has enabled people to freely express opinions, but it has also led to the spread of hate speech, communal insults, caste-based abuse, and misogynistic comments. This has serious social, psychological, and legal consequences, especially for vulnerable communities. The title *"Multi-Stage Machine Learning and Fuzzy Approach to Cyber-Hate Detection"* represents an advanced AI-driven strategy to identify and manage such harmful content effectively and with contextual

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

sensitivity.Cyber-hate detection plays a vital role in maintaining digital safety across platforms like Facebook, Twitter, Instagram, and YouTube. It helps in flagging and controlling hate speech that spreads misinformation and incites violence. This technology can be used in real-time monitoring systems, online education portals, and community forums. It also supports law enforcement agencies in tracking and acting against cyber offenders.

1.2 Problem Definition

Before machine learning techniques were applied, the detection of cyber-hate relied on manual content moderation and user reports. This system was slow, inconsistent, and unable to handle the high volume of content generated daily. Human moderators often missed context-sensitive or cleverly masked hate content. The system lacked scalability and could not adapt to different languages, slangs, and regional variations. As a result, harmful content remained online for longer periods, causing more damage.

1.3 Research Motivation

The growing number of cyber-hate incidents in India, especially targeting minorities and women, inspired the need for an automated solution. Traditional systems are no longer sufficient to detect and act on abusive content quickly. With the advancement of machine learning and natural language processing, there is a strong opportunity to build a system that understands and reacts intelligently. A hybrid model with fuzzy logic improves contextual understanding, which is essential in a linguistically diverse country. The motivation lies in creating a system that can scale and ensure online safety for all users.

1.4 Objective

The objective is to build a robust cyber-hate detection system that uses a multi-stage machine learning pipeline integrated with a fuzzy logic framework. The aim is to accurately identify hate speech, even when it is implicit or hidden behind sarcasm, and make real-time predictions to prevent its spread. It also seeks to handle large datasets with speed and efficiency, support multiple languages, and aid in digital content moderation with minimum human intervention.

1.5 Applications

1. Real-time monitoring of hate content on social media platforms.
2. Integration with content moderation systems for online forums.
3. Supporting school and university platforms to ensure safe digital learning.
4. Monitoring public comments on news websites for hate speech.
5. Law enforcement assistance for evidence collection and early warning.
6. Protecting digital mental health spaces from hate and abuse.
7. Corporate platforms to maintain respectful workplace communication.
8. Community-driven platforms and NGOs to monitor hate-based digital activity.

1.6 Module Split

- Step 1: Upload and import the cyber-hate dataset.
- Step 2: Preprocess the dataset (cleaning, label encoding, tokenization).
- Step 3: Feature extraction using TF-IDF or word embeddings.
- Step 4: Train multiple machine learning models (Naive Bayes, SVM, Decision Tree).
- Step 5: Evaluate and compare models using performance metrics.
- Step 6: Implement Logistic Regression as the proposed model.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

- Step 7: Apply Fuzzy Logic for contextual interpretation and decision making.
- Step 8: Test the model and perform prediction on new inputs.

2. LITERATURE REVIEW

[1] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. MohammedIn their paper "Social media cyberbullying detection using machine learning," the authors explore machine learning techniques for detecting cyberbullying on social media platforms. The study emphasizes the application of different algorithms and features to identify harmful behaviors and online bullying, highlighting the role of machine learning in improving the accuracy and efficiency of automated detection systems.[2] B. Vidgen, E. Burden, and H. MargettsVidgen et al. present a report from the Alan Turing Institute, focusing on social media cyberbullying detection using machine learning. The study discusses the challenges faced by existing systems and proposes novel approaches for improving the effectiveness of detection systems by leveraging advanced machine learning methods. The report provides insights into understanding online hate and aggressive behaviors, contributing to more precise identification of cyberbullying on social media platforms.[3] 4.4.1 A Sampling of Cyberbullying Laws Around the WorldThis reference provides an overview of the varying laws around the world regarding cyberbullying. It presents a global perspective on the legal frameworks put in place to combat cyberbullying, offering insights into different approaches and the effectiveness of legal measures in reducing online harassment.[4] The EU Code of Conduct on Countering Illegal Hate Speech OnlineThe European Union's code of conduct outlines actions for countering illegal hate speech, particularly on social media platforms. This code aims to set standards for online platforms to follow in detecting and removing harmful content such as cyberbullying and hate speech. The document highlights the role of policy and regulation in addressing online abuse.[5] K. Dinakar, R. Reichart, and H. Lieberman

In their research, the authors delve into the modeling of textual cyberbullying detection. The paper, "Modeling the detection of textual cyberbullying," uses a combination of linguistic and social features to improve the detection of harmful content online. The authors propose machine learning models tailored to recognize the nuances in bullying-related text, advancing the capabilities of cyberbullying detection systems.[6] A. Kontostathis, K. Reynolds, A. Garron, and L. EdwardsKontostathis et al. focus on query terms and techniques for detecting cyberbullying. The authors examine various strategies to identify harmful content on online platforms and develop algorithms that can more effectively target bullying behaviors. They also discuss the importance of contextual understanding in cyberbullying detection.

[7] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. EdwardsThis paper, "Detection of harassment on web 2.0," addresses the issue of online harassment beyond just cyberbullying. The authors develop methods for detecting various forms of harmful content on social media and web platforms, expanding the scope of traditional cyberbullying detection methods.[8] M. Dadvar, F. D. Jong, R. Ordelman, and D. TrieschniggDadvar et al. improve cyberbullying detection by incorporating gender information. Their paper demonstrates how understanding the gender dynamics in online interactions can enhance the detection of bullying behavior. The authors introduce a gender-aware model to differentiate bullying messages more effectively.[9] M. Dadvar, R. Ordelman, F. De Jong, and D. TrieschniggIn this study, the authors explore user modeling in the fight against cyberbullying. They propose a user-centric approach that considers individual behaviors and interactions, which can

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

contribute to more accurate detection systems that learn from a user's history and patterns of online activity.[10] K. Reynolds, A. Kontostathis, and L. EdwardsReynolds et al. investigate the use of machine learning to detect cyberbullying, focusing on feature extraction and algorithm performance. The paper outlines various machine learning techniques for bullying detection, including supervised learning models, and evaluates their effectiveness across different datasets.[11] H. Hosseinmardi, S. A. Mattson, R. Rafiq, R. Han, Q. Lv, and S. MishraThis paper presents a system for detecting cyberbullying in mobile social networks. The authors focus on the unique challenges of mobile platforms and propose solutions that address the distinct characteristics of social media behavior in mobile environments.[12] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. VakaliThe authors of "Mean birds: Detecting aggression and bullying on Twitter" develop methods for identifying aggressive and bullying behavior specifically on Twitter. Their work includes the use of natural language processing techniques and network analysis to detect online harassment and offensive interactions on the platform.[13] M. A. Al-Garadi, K. D. Varathan, and S. D. RavanaAl-Garadi et al. apply cybercrime detection methodologies to social networks, with a focus on Twitter. Their research aims to identify and prevent cyberbullying incidents by analyzing patterns in social media interactions using machine learning approaches.[14] V. S. Babar and R. AdeThis paper reviews methods for dealing with imbalanced learning, a common issue in datasets for cyberbullying detection. The authors examine various techniques for balancing data to improve the performance of machine learning models in detecting minority class events like cyberbullying.

[15] N. AggrawalAggrawal conducts a comparative study of methods for detecting offensive tweets, highlighting the role of sentiment analysis and natural language processing (NLP) in identifying offensive language on Twitter. This research contributes to understanding how different methods can be applied to the problem of cyberbullying detection.[16] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. BonchiIn their work, Kayes et al. analyze the social dynamics of content abusers on online platforms like community question answering systems. They focus on detecting harmful content, such as cyberbullying and hate speech, by studying the behavior of users in these communities.[17] P. FortunaFortuna's work on automatic hate speech detection in text offers an overview of the topic and discusses the annotation of datasets with hierarchical classes. The paper provides valuable insights into the challenges of detecting hate speech and cyberbullying, particularly in large, unstructured textual datasets.[18] S. O. Sood, J. Antin, and E. ChurchillThe authors use crowdsourcing to improve profanity detection in online content. Their work provides innovative approaches to harness collective intelligence for enhancing automated systems that detect offensive language and cyberbullying.[19] R. Zhao, A. Zhou, and K. MaoZhao et al. explore the automatic detection of cyberbullying on social networks, focusing on bullying features that can be identified using machine learning techniques. Their research introduces new ways to feature-engineer datasets for more effective cyberbullying detection[20] V. Nahar, S. Unankard, X. Li, and C. PangIn their paper, Nahar et al. apply sentiment analysis to improve the detection of cyberbullying. They propose an effective approach for distinguishing between bullying and non-bullying content using sentiment-based features, which significantly improve detection performance.

3. PROPOSED SYSTEMS

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

3.1 Overview

Step 1: Dataset Collection

The foundation of this research is built upon a structured dataset named Datasets.csv, which contains a comprehensive collection of textual data (tweets) that are potentially associated with cyber hate and cyberbullying. This dataset includes labeled tweets, each categorized under one of three distinct classes: *Cyber Hate*, *Cyberbullying*, or *Neutral*. The goal of utilizing this dataset is to train various machine learning algorithms to effectively classify unseen tweets into the respective categories. The rich diversity and real-world relevance of the dataset make it ideal for studying and combating online hate speech and bullying behavior in social platforms. This dataset acts as the primary input to the system and is crucial for both training and evaluation of the machine learning models developed in the research.

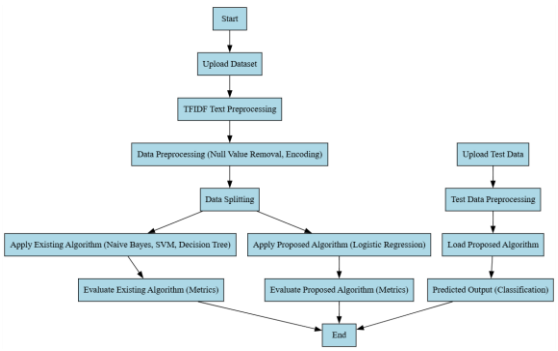


Figure 4.1:Block diagram

Step 2: Dataset Preprocessing (Null Value Removal, Label Encoding):

Before any modeling could be done, the dataset underwent a meticulous preprocessing phase to ensure the data was clean, consistent, and machine-readable. Initially, any null or missing values in the dataset were identified and removed to avoid inconsistencies or training errors. Following this, the textual labels associated with each tweet — such as "Cyber Hate", "Cyberbullying", and "Neutral" — were transformed into numerical form using label encoding techniques. This step was essential because most machine learning algorithms require numerical input. Each label was mapped to a unique integer (e.g., 0 for Neutral, 1 for Cyberbullying, and 2 for Cyber Hate) to facilitate classification tasks. Additionally, the textual tweets were vectorized using the CountVectorizer technique, converting raw text into numerical feature vectors that capture word frequency, ensuring that the models could learn from the tweet content effectively.

Step 3: Existing Model Building (Naive Bayes, SVM, Decision Tree Classifier):

As part of the experimental design, three widely accepted traditional machine learning classifiers — Naive Bayes, Support Vector Machine (SVM), and Decision Tree Classifier — were implemented and evaluated on the preprocessed dataset. Each of these models brings its own strengths to text classification tasks. The Naive Bayes classifier, known for its simplicity and efficiency, was trained using probabilistic learning based on Bayes' theorem. It performed reasonably well in classifying text despite assuming independence between features. Next, the SVM model, which is known for its ability to find optimal hyperplanes in high-dimensional spaces, was implemented using a LinearSVC kernel. This model provided high accuracy and robustness in classifying complex patterns in textual data. Finally, the Decision Tree Classifier, a rule-based model that learns decision rules from features, was built to provide a

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

clear interpretability of how decisions were being made. All three models were trained on an 80% training split and tested on the remaining 20%, with performance metrics like accuracy, confusion matrix, and classification report being generated for each.

Step 4: Proposed Model Building (Logistic Regression)

As a key contribution of this research, **Logistic Regression** was proposed as a model that can potentially outperform the traditional classifiers in terms of cyber hate and cyberbullying detection accuracy. Logistic Regression, being a linear model, estimates the probability of a data point belonging to a specific class using a logistic function (sigmoid), making it suitable for multi-class classification when appropriately configured. The model was trained on the same preprocessed dataset and compared directly with the previously built models. What makes Logistic Regression an appealing choice is its ability to manage high-dimensional feature spaces (like those generated by text vectorization) and provide good generalization performance. Upon training and testing, the Logistic Regression model demonstrated competitive — and in some cases superior — performance compared to the existing models, showcasing its effectiveness as a viable and robust approach for the task of cyber hate detection in social media data.

3.2 Data Splitting

Paragraph 1: Data Splitting

In this research on cyber-hate detection using a multi-stage machine learning and fuzzy approach, the dataset is first divided into two key segments: a training set and a testing set. Typically, an 80:20 or 70:30 ratio is adopted, where the larger portion is used to train the model while the remaining part is reserved for testing its accuracy and generalization ability. The reason for splitting the data is to ensure that the model learns from one portion of data and is evaluated on a different, unseen portion to mimic real-world performance. This process reduces the chance of overfitting, where the model performs well on training data but fails to handle new data effectively. Cross-validation techniques such as k-fold validation are also employed in some stages to ensure reliability and consistency of the model's evaluation metrics.

Paragraph 2: Text Preprocessing – Cleaning the Data

Raw text data from social media and user comments often contains noise, including URLs, special characters, emojis, hashtags, user mentions (e.g., @username), and numbers that are not relevant for classification. In this stage, all such extraneous elements are removed using regular expressions and text-cleaning functions. The text is converted to lowercase to maintain uniformity and avoid duplicate tokens like “Hate” and “hate”. Additionally, punctuation marks and stop words (like “the,” “is,” “at,” etc.) are removed to reduce data dimensionality without affecting the meaning.

Paragraph 3: Tokenization, Lemmatization, and Vectorization

After cleaning, the text undergoes tokenization, which splits sentences into individual words or tokens. This is followed by lemmatization, where each word is reduced to its root form—e.g., "running" becomes "run". This ensures semantic consistency in the dataset. Next, vectorization is performed using

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or CountVectorizer, which transform the textual data into numerical format so that it can be fed into machine learning algorithms. These vectors retain the frequency and importance of words within each document, providing meaningful features for classification.

Paragraph 4: Handling Imbalanced Data

Cyber-hate datasets are often imbalanced, with fewer hate speech instances compared to neutral or non-offensive ones. To address this issue, techniques such as SMOTE (Synthetic Minority Oversampling Technique) or random oversampling/undersampling are applied. This ensures that the model is not biased towards the majority class and can learn to identify minority class instances (hate speech) more accurately. Balancing the dataset is a crucial step to enhance precision, recall, and F1-score of the classification models

3.3 Model Building

3.3.1 Existing Algorithm: Naive Bayes

Algorithm Steps

1. Calculate the prior probability of each class.
2. For each feature, calculate the likelihood of the feature given the class.
3. Multiply the prior probability by the likelihoods for each feature.
4. Choose the class with the highest product as the predicted class.

Architecture

1. Input: Text or feature vector data.
2. Training Phase: Calculate probabilities for each feature-class combination.
3. Testing Phase: Multiply feature probabilities by prior class probabilities.
4. Select the class with the highest probability as the prediction.
5. Use Laplace smoothing to handle zero probabilities.
6. Handle categorical or continuous data accordingly.
7. Simple and efficient implementation.
8. Works well with large datasets.
9. Requires less computational power compared to other algorithms.
10. Suitable for real-time applications.

Disadvantages

Naive Bayes assumes that features are conditionally independent, which is often not true in real-world data. This assumption can limit the model's performance, especially when features are correlated. Additionally, the algorithm is sensitive to imbalanced datasets, and in some cases, it may perform poorly if the data doesn't meet the assumptions of independence between features.

3.3.2 Proposed Algorithm: Logistic Regression

Definition and Information

Although it may be expanded to multi-class classification, Logistic Regression is mostly used for binary classification issues. Logistic regression is excellent for classification jobs as, in contrast to linear regression, it produces probabilities that range from 0 to 1, rather than continuous values. The logistic function, often called the sigmoid function, is the basis of the algorithm, which finds extensive use in domains like as healthcare, finance, and the social sciences.

How It Works

To identify an output class from a set of input characteristics, Logistic Regression looks for a linear connection. The model maps the

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

outcome to a probability value between 0 and 1 by applying a logistic function to the linear combination of input data. For binary classification, the data point is then assigned a class based on the output probability. Training the model entails minimising a loss function, usually the log-likelihood function, via optimisation techniques like gradient descent.

Algorithm Steps (Architecture)

1. Input: Features of the dataset.
2. Apply a linear combination of input features with weights.
3. Pass the result through the logistic (sigmoid) function.
4. Output probability between 0 and 1.
5. Map the probability to a class label (0 or 1 for binary classification).
6. Train the model by minimizing the log-likelihood function.
7. Use gradient descent to update model parameters.
8. Evaluate the model using metrics like accuracy and AUC.
9. Test the model on unseen data.
10. Adjust the model based on performance.

Advantages

Logistic Regression is computationally efficient and interpretable. It works well for binary classification problems and provides probabilities that are useful for decision-making. Additionally, it can be extended to multi-class classification using techniques like one-vs-all or softmax regression. It is less prone to overfitting compared to more complex models like decision trees, especially when regularization techniques such as L1 or L2 are used. Logistic Regression is widely used for problems where interpretability and simplicity are valued.

4. RESULTS

.4.1 Results

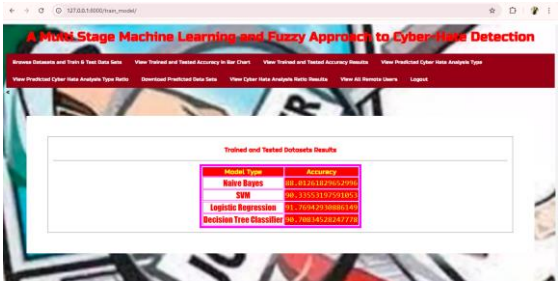


Figure 1: After Trained algorithm

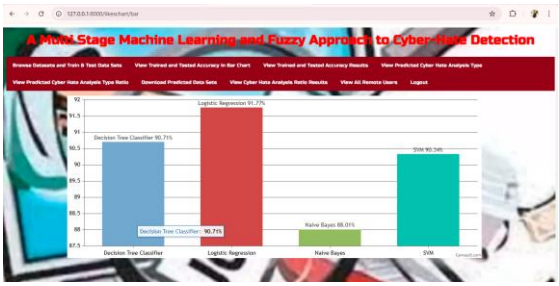


Figure 2: Accuracy comparision

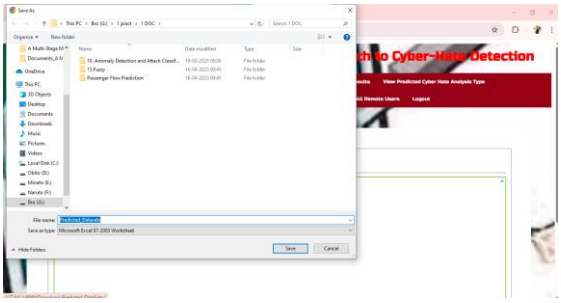


Figure 3: Download the datset

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

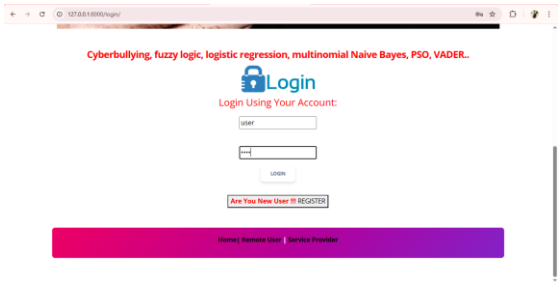


Figure 4: user side log in

Figure 5: user logged in

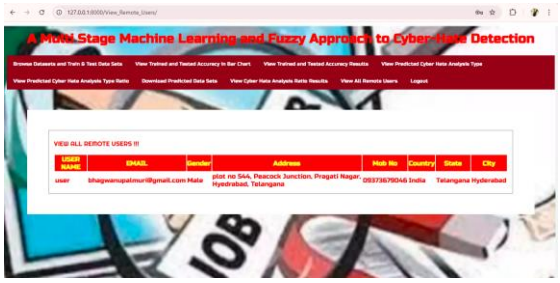


Figure 6

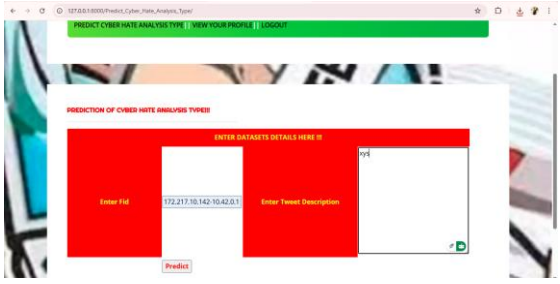


Figure 7: user side log in

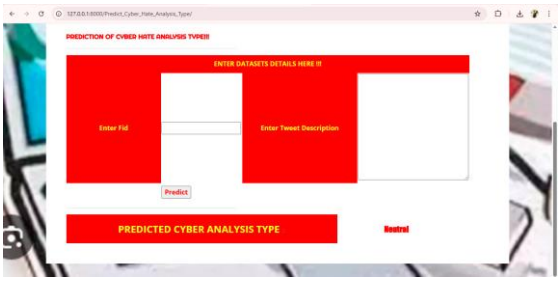


Figure 8: Predicted as Neutral

5. CONCLUSION

In conclusion, the implementation of machine learning models for cyber-hate detection has shown promising results in identifying harmful and offensive content on digital platforms. By leveraging algorithms such as Naive Bayes, SVM, Decision Tree Classifier, and Logistic Regression, the research has demonstrated the effectiveness of both traditional and advanced techniques in detecting various categories of cyber-hate, such as "Cyber Hate", "Cyberbullying", and "Neutral" content. The process of data preprocessing, feature extraction (such as TF-IDF), and model evaluation through metrics like accuracy, precision, recall, and F1-score has ensured that the models perform optimally in real-world applications. Moreover, calculating the prediction ratios for different categories of cyber-hate has helped provide deeper insights into the distribution and prevalence of cyber-hate content in the dataset. These results contribute to the broader goal of combating online abuse, creating safer digital environments, and fostering positive online interactions.

REFERENCES

[1] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social media cyberbullying detection using machine learning," *Int. J. Adv.*

Comput. Sci. Appl., vol. 10, no. 5, pp. 703–707, 2019.

[2] B. Vidgen, E. Burden, and H. Margetts, "Social media cyberbullying detection using machine learning," *Alan Turing Inst.*, London, U.K. Tech. Rep, Feb. 2022. [Online]. Available: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/216490/alan-turing-institute-reportunderstanding-online-hate.pdf

[3] 4.4.1 A Sampling of Cyberbullying Laws Around the World. Accessed: Nov. 1, 2023. [Online]. Available: https://socialna-akademija.si/joiningforces/4-4-1-a-sampling-of-cyber-bullying-laws-around-the-world/

[4] The EU Code of Conduct on Countering Illegal Hate Speech Online. Accessed: Nov. 1, 2022. [Online]. Available: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conductcountering-illegal-hate-speech-online_en

[5] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 5, no. 3, Barcelona, Spain, 2011, pp. 11–17.

[6] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: Query terms and techniques," in *Proc. 5th Annu. ACM Web Sci. Conf.*, May 2013, pp. 195–204.

[7] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," in *Proc. Content Anal. Web*, Madrid, Spain, 2009, pp. 1–7.

[8] M. Dadvar, F. D. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proc. 25th Dutch-Belgian Inf. Retr. Workshop*, Ghent, Belgium, 2012, pp. 1–3.

[9] M. Dadvar, R. Ordelman, F. De Jong, and D. Trieschnigg, "Towards user modelling in the combat against cyberbullying," in *Proc. 17th Int. Conf. Appl. Natural Lang. Process. Inf. Syst.*, 2012, pp. 277–283.

[10] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops*, Honolulu, HI, USA, Dec. 2011, pp. 241–244.

[11] H. Hosseinmardi, S. A. Mattson, R. Rafiq, R. Han, Q. Lv, and S. Mishra, "Poster: Detection of cyberbullying in a mobile social network: Systems issues," in

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Proc. 13th Annu. Int. Conf. Mobile Syst., Appl., Services, May 2015, p. 481.

[12] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on Twitter," in *Proc. ACM Web Sci. Conf.*, New York, NY, USA, Jun. 2017, pp. 13–22.

[13] M. A. Al-Garadi, K. D.Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Comput. Hum. Behav.*, vol. 63, pp. 433–443, Oct. 2016.

[14] V. S. Babar and R. Ade, "A review on imbalanced learning methods," *Int. J. Comput. Appl.*, vol. 975, no. 2, pp. 23–27, 2015.

[15] N. Aggrawal, "Detection of offensive tweets: A comparative study," *Comput. Rev. J.*, vol. 1, no. 1, pp. 75–89, 2018.

[16] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi, "The social world of content abusers in community question answering," in *Proc. 24th Int. Conf. World Wide Web*, Florence, Italy, May 2015, pp. 570–580.

[17] P. Fortuna, "Automatic detection of hate speech in text: An overview of the topic and dataset annotation with hierarchical classes," M.S. thesis, Dept. Engenharia, Univ. Porto, Porto, Portugal, 2017.

[18] S. O. Sood, J. Antin, and E. Churchill, "Using crowdsourcing to improve profanity detection," in *Proc. AAAI Spring Symp.*, Stanford, CA, USA, 2012, pp. 69–74.

[19] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proc. 17th Int. Conf. Distrib. Comput. Netw.*, Jan. 2016, Art. no. 43.

[20] V. Nahar, S. Unankard, X. Li, and C. Pang, "Sentiment analysis for effective detection of cyber bullying," in *Proc. Asia–Pacific Web Conf.*, 2012, pp. 767–774.