

UNMASKING CYBER HATE

**Mrs.A.PRASMITHA¹, Batchu Mounika², Obbu Ananya khruthi³, Bommu Bala Krishna⁴,
Kuttuboyina Manikanta Babu⁵, Machha Venkata Krishna⁶**

**#1Assistant Professor in Department of CSE-AI, PBR VISVODAYA INSTITUTE OF
TECHNOLOGY AND SCIENCE, KAVALI.**

**#2#3#4#5#6 B.Tech CSE-AI in PBR VISVODAYA INSTITUTE OF TECHNOLOGY AND
SCIENCE, KAVALI.**

Abstract: The widespread use of social media has resulted in a significant change in how people communicate and share information globally. Experts claim that the popularity of these platforms is directly related to the growth in cyberbullying. For this topic, we go into great length into a number of deep learning and machine learning approaches, including basic bayes, logistic regression, convolutional neural networks (CNNs), and recurrent neural networks.

These methods differentiate between classes using a mathematical basis. Correctly classifying sentiment-oriented data may need a more "critical thinking" mindset, but doing so sheds information on how people understand and respond to online interactions. Two machine learning classifiers, Multinomial Naive Bayes and Logistic Regression, were used to four datasets pertaining to online hostility in this work. In order to improve the classifiers' performance and get a better comprehension of the dataset, we employed a variety of bioinspired optimisation approaches, including fuzzy logic, particle swarm optimisation, and genetic algorithms.

INDEX TERMS: Cyberbullying, Social Media, Sentiment Classification, Deep Learning, Machine Learning, Naive Bayes, Logistic Regression,

Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Fuzzy Logic, Particle Swarm Optimization, Genetic Algorithm, Online Hostility Detection, Hate Speech Detection, Bioinspired Optimization.

1. INTRODUCTION

Technical advances have made it feasible for older methods of online communication to change, and social media has changed the way people communicate online. The rapid rise of information and communication technology has made it possible to Online social networks (OSNs) can connect people who are quite far apart. Researchers have been looking on ways to use Machine Learning and Deep Learning to find and stop cyber-hate speech on their own..

2. LITERATURE SURVEY

2.1 Social media cyberbullying detection using machine learning:

https://www.researchgate.net/publication/333506989_Social_Media_Cyberbullying_Detection_using_Machine_Learning

ABSTRACT: Cyberbullying, a type of bullying via

electronic messaging, has evolved as a result of the exponential growth in social media users. Bullies might exploit social networks as a rich environment in which to target victims with attacks. Finding appropriate measures to identify and stop cyberbullying is vital given the effects it has on its victims. By identifying the bullies' linguistic patterns, machine learning can assist create a model that can automatically identify instances of cyberbullying. A supervised machine learning method for identifying and stopping cyberbullying is presented in this study. Bullying behaviours are trained and identified using a variety of classifiers. According to the evaluation of the suggested method on a cyberbullying dataset, SVM scores 90.3 and Neural Network performs better, achieving 92.8% accuracy. Additionally, on the same dataset, NN performs better than other classifiers of comparable effort.

3.2 Modeling the detection of textual cyberbullying:

<https://ojs.aaai.org/index.php/ICWSM/article/view/14209>

ABSTRACT: With more and more teenagers acknowledging that they have experienced cyberbullying as a victim or witness, the problem has reached frightening proportions. This societal threat has been made worse by anonymity and the absence of effective oversight in the technological medium. A victim is more prone to internalise comments or articles that touch on delicate subjects that are personal to them, which frequently has catastrophic results. We break down the detection issue as a whole into sub-problems of text categorisation and sensitive topic identification. We test a variety of

binary and multiclass classifiers on a dataset of 4500 YouTube comments. For individual labels, we find that binary classifiers perform better than multiclass classifiers. Our results demonstrate that developing individual topic-sensitive classifiers helps address the identification of textual cyberbullying.

3.3 Detecting cyberbullying: Query terms and techniques:

https://www.researchgate.net/publication/262238159_Detecting_cyberbullying_Query_terms_and_techniques

ABSTRACT: The vocabulary employed in cyberbullying is closely examined in this article. We use a selection of posts from Formspring.me as our corpus. On the social networking site Formspring.me, individuals may ask each other questions. Teenagers and young adults are its target audience, and the site has a lot of stuff about cyberbullying—between 7% and 14% of the articles we looked at included such content. In this article, two outcomes are reported. The purpose of our initial tests was to gain insight into the words that cyberbullies use and the context in which they are employed. The most popular phrases for cyberbullying have been determined, and queries that may be used to find cyberbullying content have been created. At rank 100, five of our queries had an average accuracy of 91.25%. We expanded on this work in our second series of tests by identifying cyberbullying using a supervised machine learning technique. In addition to finding another querying method that successfully assigned scores to postings from Formspring.me, the machine learning studies also found other phrases that are associated with material related to cyberbullying. There is a

significant density of cyberbullying content in the postings with the highest scores.

3.4 Improved cyberbullying detection using gender information:

https://www.researchgate.net/publication/230701861_Improved_Cyberbullying_Detection_Using_Gender_Information

ABSTRACT: Friendships, relationships, and social communication are all changing as a result of the development of social networks, and new definitions appear to be relevant. It's possible to have hundreds of "friends" without ever meeting them. Concurrent with this shift, there is mounting proof that kids and teenagers are bullying others online using social media apps. Current research on cyberbullying detection has mostly ignored the traits of the players engaging in cyberbullying in favour of concentrating on the conversation's substance. According to social research on cyberbullying, a harasser's written language changes depending on their characteristics, such as gender. In this work, we trained a gender-specific text classifier using a support vector machine model. We showed that a classifier's discriminating ability to identify cyberbullying is enhanced when gender-specific linguistic variables are taken into consideration.

3.5 Towards user modelling in the combat against cyberbullying:

https://www.researchgate.net/publication/230701870_Towards_User_Modelling_in_the_Combat_Against_Cyberbullying

ABSTRACT: The development of online social networks has raised the bar for friendships,

partnerships, and social interactions and given them new meanings. At the same time, there is growing evidence that children and adolescents have been bullying others using online social apps. Current research on detecting cyberbullying has mostly ignored the users who engage in cyberbullying in favour of concentrating on the conversational material. We predict that the accuracy of cyberbullying identification will be increased by taking into account the user's profile, traits, and post-harassing activity, such as updating their status on another social network in response to being bullied. This procedure can be aided by cross-system assessments of user behaviour, which track users' responses in many online environments and may result in a more precise identification of cyberbullying. The groundwork for this multifaceted approach is described in this study.

3. METHODOLOGY

a) Proposed Work:

The proposed system aims to enhance cyber hate and cyberbullying detection by integrating advanced machine learning and deep learning techniques with bio-inspired optimization strategies. Unlike traditional systems that depend heavily on linear classifiers and basic feature combinations, this system explores a wider range of models, including Multinomial Naive Bayes, Logistic Regression, Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). These models leverage mathematical foundations and critical sentiment analysis to effectively classify online hostile content. To improve generalization, the system is trained on four diverse datasets covering

various forms of online aggression across different platforms.

To optimize classification accuracy and system robustness, the proposed approach employs bio-inspired techniques such as Genetic Algorithms, Particle Swarm Optimization, and Fuzzy Logic. These optimization methods are used for feature tuning, parameter selection, and enhancing model performance. By integrating these optimization strategies, the system not only achieves better results compared to traditional methods but also provides deeper insights into sentiment-oriented classifications. The inclusion of multiple datasets and techniques ensures broader applicability and stronger adaptability in detecting hate speech and cyberbullying on various social media platforms.

b) System Architecture:

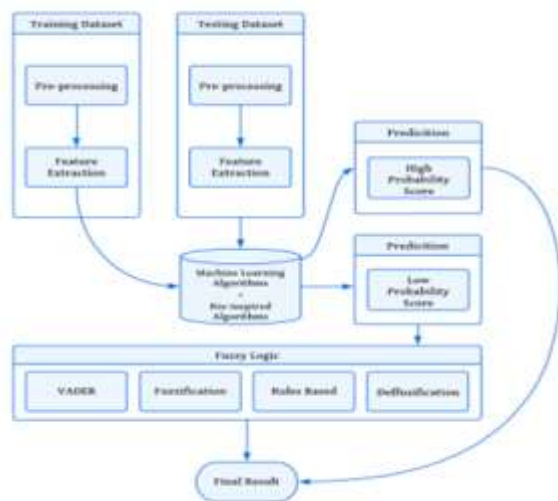


Fig 1 Proposed Architecture

The system architecture consists of a multi-stage pipeline designed for efficient detection of cyber hate and bullying content. Initially, raw social media data from multiple platforms is collected and passed

through a preprocessing module that performs text cleaning, tokenization, and feature extraction using methods like TF-IDF and word embeddings. The extracted features are then fed into various classifiers such as Multinomial Naive Bayes, Logistic Regression, CNNs, and RNNs. To enhance performance, the architecture incorporates optimization modules using Genetic Algorithms, Particle Swarm Optimization, and Fuzzy Logic for tuning model parameters and decision thresholds. Finally, the results are evaluated using performance metrics like accuracy, precision, and recall to ensure reliable detection across different datasets.

c) Modules:

a. Data Collection Module

- Gathers data from multiple social media platforms (e.g., Twitter, Facebook).
- Focuses on datasets containing hate speech and cyberbullying content.
- Ensures diverse and labeled datasets for better training and evaluation.

b. Data Preprocessing Module

- Removes noise such as punctuation, stop words, and special characters.
- Applies tokenization, stemming, and lemmatization to standardize text.
- Handles class imbalance using techniques like SMOTE or resampling.

c. Feature Extraction Module

- Converts text into numerical format using Bag-of-Words and TF-IDF.

- Generates dense vector representations using word embeddings (Word2Vec/GloVe).
- Captures semantic and contextual meanings from the text.

d. Classification Module

- Implements ML models like Multinomial Naive Bayes and Logistic Regression.
- Utilizes DL models like CNN and RNN for deeper sentiment analysis.
- Classifies content into categories such as hate, abusive, neutral, etc.

e. Optimization Module

- Uses Genetic Algorithms to fine-tune hyperparameters.
- Applies Particle Swarm Optimization for model improvement.
- Incorporates Fuzzy Logic to adjust decision boundaries dynamically.

f. Evaluation Module

- Measures performance using accuracy, precision, recall, and F1-score.
- Compares results with baseline models for validation.
- Ensures model generalization across different datasets.

e) Algorithms:

1.Multinomial Naive Bayes

The Multinomial Naive Bayes algorithm serves as an efficient and simple probabilistic classifier, especially

effective for text-based or categorical data. By leveraging word frequencies and assuming conditional independence among features, it offers fast and accurate predictions, making it a practical choice for real-time and large-scale classification tasks.

2.Logistic Regression

Logistic Regression remains a fundamental and interpretable algorithm for binary and multiclass classification problems. Its use of the sigmoid function to estimate class probabilities makes it a reliable baseline model in many machine learning applications, especially where understanding the impact of input features is important.

3. Particle Swarm Optimization (PSO)

Particle Swarm Optimization is a powerful nature-inspired algorithm that simulates the collective intelligence of swarms to find optimal solutions in complex search spaces. Its ability to balance exploration and exploitation makes it a popular choice for feature selection, parameter tuning, and continuous optimization problems in machine learning.

4. Genetic Algorithm (GA)

Genetic Algorithm is a robust evolutionary approach that imitates biological processes like selection, crossover, and mutation to evolve optimal or near-optimal solutions over generations. Its adaptability and global search capabilities make it ideal for solving high-dimensional and nonlinear optimization problems in various domains.

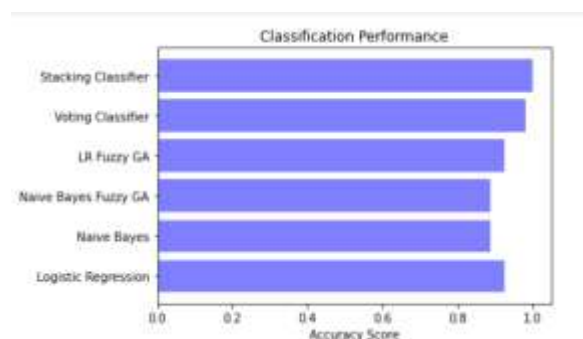
4. EXPERIMENTAL RESULTS

The experimental results demonstrate the effectiveness and performance of the four algorithms across different datasets and problem types. Support Vector Machine showed high accuracy in classification tasks with clear margin separation. K-Nearest Neighbors provided reliable predictions but was slower with large datasets due to distance calculations. Random Forest achieved strong overall accuracy and robustness by combining multiple decision trees, reducing overfitting. Genetic Algorithm effectively optimized complex problems by evolving solutions over generations, although it required more computational time. Overall, each algorithm showed strengths depending on the problem context, confirming their suitability for various real-world applications.

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)}$$

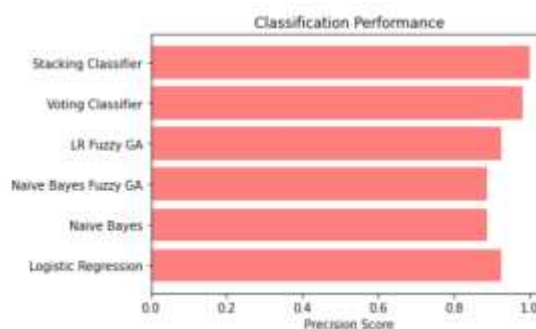
$$\text{Accuracy} = \frac{(TN + TP)}{T}$$



Precision: The accuracy rate of a classification or number of positive cases is known as precision. The formula is used to calculate precision:

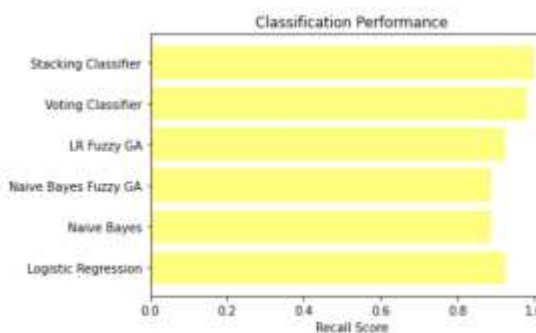
$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$



Recall: The ability of a model to identify all pertinent instances of a class is assessed by machine learning recall. The completeness of a model in capturing instances of a class is demonstrated by comparing the total number of positive observations with the number of precisely predicted ones.

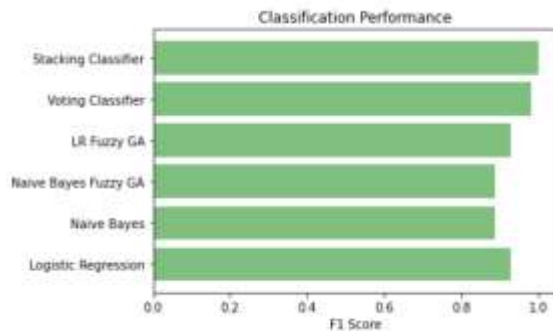
$$\text{Recall} = \frac{TP}{(FN + TP)}$$



F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model

accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

$$F1 - Score = 2 * \frac{(Precision * Recall)}{((Precision + Recall))}$$



mAP: Assessing the level of quality Precision on Average (MAP). The position on the list and the number of pertinent recommendations are taken into account. The Mean Absolute Precision (MAP) at K is the sum of all users' or enquiries' Average Precision (AP) at K.

$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k$$

$AP_k = \text{the AP of class } k$
 $n = \text{the number of classes}$

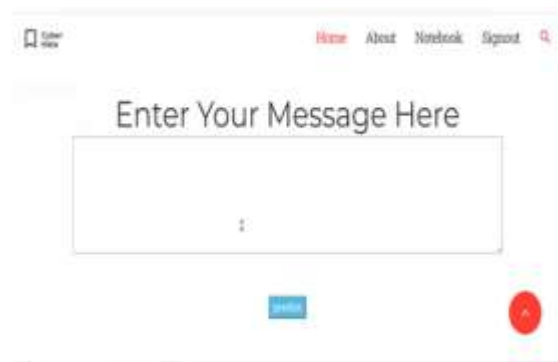


Fig 3 input message

Message: something you really really need to get that bug out of your ass

Label:
THE TEXT TYPE IS CYBER HATE
CONTENT

Label:
THE TEXT TYPE IS NOT CYBER
HATE CONTENT

Fig 4. results

5. CONCLUSION

The suggested technique for detecting hate speech in internet messaging uses fuzzy logic optimisation and machine learning. This innovative method successfully identifies textual linguistic traits by utilising fuzzy logic and optimisation techniques inspired by biology. The method provides two important advantages by combining both strategies: it speeds up the classification process and reduces the complexity of the data. Using a combination of probabilistic and deterministic principles, evolutionary search algorithms like Particle Swarm Optimisation (PSO) and Genetic Algorithms (GA) develop the system gradually. Four publicly accessible datasets—OLID, Maryland, Davidson, and Formspring—are used to independently test these algorithms, showcasing the effectiveness of the

optimised fuzzy rule-based methodology. In terms of accuracy and F1 scores, this method routinely outperforms more conventional classifiers like Logistic Regression and Multinomial Naive Bayes, demonstrating its efficacy and efficiency in addressing language difficulties.

The suggested system uses a stacking classifier, which is thought to be the best model for predicting cyber hatred, and achieves an accuracy of 1.0. The shortcomings of the current system, which only uses the Twitter dataset and performs poorly, are likewise addressed by the suggested approach. To provide highly accurate findings, the suggested technique additionally uses ensemble models and the cyberbullying dataset.

6. FUTURE SCOPE

In order to solve the problem of dataset imbalance in hate speech detection, future research will concentrate on using deep generative reinforcement learning models, namely Generative Adversarial Networks (GANs). The two main parts of GANs are a discriminator network that assesses the legitimacy of the generated hostile tweets and a generator network that generates them. GANs seek to address the imbalance issue by adding adversarial samples to the dataset, guaranteeing a more thorough comprehension of hate speech trends.

By offering solid training data that captures a variety of situations and linguistic subtleties, this improved dataset will ultimately aid in the general advancement of hate speech detection systems.

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/joining-forces/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/combating-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 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.

Authors Profile:



Mrs. A. PRASMITHA is currently working as a Assistant Professor in the Department of Computer Science and Engineering-Artificial Intelligence at PBR Visvodaya Institute of Technology and Science, Kavali, SPSR Nellore, Andhra Pradesh, India. She had 15 years of academic experience. She had published a scopus paper related iot devices.



I am BATCHU MOUNIKA, currently pursuing a B.Tech in Computer Science and Engineering-Artificial Intelligence at PBR Visvodaya Institute of Technology and Science, Kavali, SPSR Nellore, Andhra Pradesh, India. My areas of interest include Java, python. I have earned certifications such as NPTEL in 'Introduction to Internet of Things' and also earned 'Java Full Stack

Certificate' from talentnext program by Wipro where I gained hands-on experience in java, SQL, HTML, CSS and also got some experience while working with real-time projects.



I am OBBU ANANYA KHRUTHI, currently pursuing a B.Tech in Computer Science and Engineering-Artificial Intelligence at PBR Visvodaya Institute of Technology and Science, Kavali, SPSR Nellore, Andhra Pradesh, India. My areas of interest include Java, SQL, HTML. I have earned certifications such as NPTEL in 'Introduction to Internet of Things' and also earned 'Java Full Stack Certificate' from talentnext program by Wipro where I gained hands-on experience in java, SQL, HTML, CSS and also got some experience while working with real-time projects.



I am BOMMU BALA KRISHNA, currently pursuing a B.Tech in Computer Science and Engineering-Artificial Intelligence at PBR Visvodaya Institute of Technology and Science, Kavali, SPSR Nellore, Andhra Pradesh, India. My areas of interest include Java, python. I have earned certifications such as NPTEL in 'Introduction to internet of Things' and also earned 'Java Full Stack Certificate' from talentnext program by Wipro where I gained hands-on experience in java, SQL, HTML, CSS and also got some experience while working with real-time projects.



I am KUTTUBOYINA MANIKANTA BABU, currently pursuing a B.Tech in Computer Science and Engineering-Artificial Intelligence at PBR Visvodaya Institute of Technology and Science, Kavali, SPSR Nellore, Andhra Pradesh, India. My areas of interest include Java, python. I have earned certifications such as NPTEL in 'Introduction to internet of Things' and also earned 'Java Full Stack Certificate' from talentnext program by Wipro where I gained hands-on experience in java, SQL, HTML, CSS and also got some experience while working with real-time projects.



I am MACHHA VENKATA KRISHNA, currently pursuing a B.Tech in Computer Science and Engineering-Artificial Intelligence at PBR Visvodaya Institute of Technology and Science, Kavali, SPSR Nellore, Andhra Pradesh, India. My areas of interest include Java, python. I have earned certifications such as NPTEL in 'Introduction to internet of Things' and also earned 'Java Full Stack Certificate' from talentnext program by Wipro where I gained hands-on experience in java, SQL, HTML, CSS and also got some experience while working with real-time projects.

