Detecting Offensive Language in Social Media Using Text Classification Techniques

Dr. Konda Hari Krishna[1] , Ms.Vankadari Yasaswini[2]

[1] M.Tech, Ph.D. Associate Professor, Dept. of CSE, khk396@gmail.com Orcid Id: 0000-0002-0244-7055, School of Computing, Mohan Babu University, Tirupati, A.P-517102. India

[2] Post Graduation of MCA, School of Computing, Mohan Babu University, vankadariyasaswini07@gmail.com, Tirupati, India

Abstract *– The increasing prevalence of offensive language on social media platforms poses a significant threat to individuals and communities, often leading to bullying and emotional harm. To address this issue, the research community has explored various supervised learning approaches and developed specialized datasets aimed at the automatic detection of offensive content. In this study, we propose a robust model for offensive language detection that integrates a modular preprocessing phase, three embedding techniques, and eight classifiers.Our model begins with a comprehensive cleaning and tokenization process to prepare the data for analysis. We then explore three different embedding methods, including Term Frequency-Inverse Document Frequency (TF-IDF), to capture the textual features effectively. The classification phase involves eight machine learning algorithms, with a focus on maximizing detection accuracy through hyperparameter optimization.The model was evaluated using a dataset collected from Twitter, a popular social media platform known for its diverse and often volatile user-generated content. Our experiments demonstrate that the combination of AdaBoost, Support Vector Machines (SVM), KNN, CNN classifiers with the TF-IDF embedding method achieved the highest average F1-scores, indicating superior performance in detecting offensive language..*

Keywords: *Offensive Language Detection, Social Media, Machine Learning, Text Classification, TF-IDF, Embedding Techniques, Support Vector Machine(SVM), AdaBoost, Text Preprocessing.*

I. INTRODUCTION

Deep learning has changed numerous scientific fields in the previous decade, including natural language processing (NLP), medical imaging healthcare, cyber security, social computing and a variety of other topics In recent years, with the seeming growth of social media, new concerns regarding users mental and physical safety have been introduced.

Based on a report in among students between 12 to 18 who reported being bullied at school, 15% were bullied online through social medias. In addition, the percentage of individuals who have experienced being victims of cyberbullying during their lifetime has more than doubled from 2017 to 2019 from 18% to 37% Offensive, hateful or threatening speech on the content exchanged by the crowd might range from minor or implicit bullying to severe and explicit violent threatening over victims with specific characteristics such as race, sex, religion, community, etc.

Shows that the rise of public media cyberbully poses a global problem that might damage people’s online lives. The state-of-the-art approaches target various contexts, domains, platforms for detecting a specific category of offensive language, e.g., hate speech with or without considering the severity. In this regard, various datasets also have been published to evaluate the correctness and precision of proposed

II. LITERATURE SURVEY

Skin 1.Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed, Author:Kogilavani Shanmugavadivel, V E Sathishkumar, Sandhiya Raja, T Bheema Lingaiah, S Neelakandan, and Malliga Subramanian

It faces the difficulties of handling code-mixed data because of noise and an absence of resources when compared to monolingual datasets. The suggested system employs machine learning and deep learning techniques, featuring pre-trained models such as BERT and RoBERTa, attaining significant accuracy levels of 65% for sentiment detection and 79% for identifying offensive language. The research highlights the significance of efficient preprocessing and the application of embeddings to improve model performance.

2.Offensive Language Detection Using Text Classification,

Author:Anas Ahmed Raheeq and Mrs. Afroze Begum

The suggested model consists of a modular cleaning stage, several embedding techniques, and different classifiers, yielding encouraging outcomes in identifying offensive language. The research highlights the significance of upholding community standards and enhancing user well-being by promptly recognizing and dealing with inappropriate content. It additionally evaluates existing algorithms and potential pathways for improving detection techniques.

3.OffensEval 2023: Offensive language identification in the age of Large Language Models, Author:Marcos Zampieri, S Rosenthal, Preslav Nakov, Alphaeus Dmonte, and Tharindu Ranasinghe

It explores the OffensEval shared tasks, which have greatly enhanced the identification of offensive language via benchmark datasets in several languages, such as Arabic, Danish, English, Greek, and Turkish. It emphasizes the application of Large Language Models (LLMs) for zero-shot prompting and contrasts their effectiveness with leading teams from OffensEval competitions. The article highlights the OLID hierarchical taxonomy as a benchmark for research on offensive language and points out the constraints of LLMs in assisting non-English languages.

4.Cross-lingual Offensive Language Detection: A Systematic Review of Datasets, Transfer Approaches and Challenges. Authors:Aiqi Jiang and A Zubiaga

The research article offers a comprehensive review of Cross-Lingual Transfer Learning (CLTL) methods for identifying offensive language in social media, evaluating 67 pertinent studies. It classifies datasets and resources, emphasizing the obstacles of language diversity, limited datasets, and model adaptability. The document highlights three primary CLTL transfer methods: instance, feature, and parameter transfer. It stresses the importance of creating new datasets for low-resource languages and developing stronger CLTL techniques. The survey acts as a benchmark for existing practices and informs future studies in cross-lingual offensive language identification.

5.Offensive Language Detection Using Soft Voting Ensemble Model, Authors:Brillian Fieri and Derwin Suhartono

The research shows that the soft voting classifier surpasses conventional models, obtaining superior F1 scores on both the original and augmented datasets. The dataset includes tweets categorized as hate speech and offensive language, featuring 20,620 offensive instances and 4,163 non-offensive ones.

6.A survey on multi-lingual offensive language detection, Authors:Khouloud Mnassri, Reza Farahbakhsh, Razieh Chalehchaleh

It offers an in-depth examination of current methods, datasets, and resources, while tackling the issues encountered in this area, such as technical, cultural, and linguistic constraints. The authors suggest future avenues for improving detection methods, highlighting the significance of collaborative initiatives and sophisticated machine learning strategies, including unsupervised learning, meta-learning, and multitask learning.

7.Investigating Offensive Language Detection in a Low-Resource Setting with a Robustness Perspective, Authors:Zuchao Li, Min Peng, Israe Abdellaoui, Anass Ibrahimi, Mohamed Amine El Bouni, Asmaa Mourhir, Saad Driouech, and Mohamed Aghzal

It emphasizes identifying offensive language in Moroccan Darija, a low-resource Arabic dialect, underscoring difficulties like the absence of standardized writing and common code-switching. A dataset labeled by humans consisting of 20,402 sentences was compiled from social media, highlighting a class imbalance with 37.8% of the content being offensive. Different BERT-based models were refined, resulting in a Darija-specific model that reached 90% accuracy. The research highlights the importance of testing for robustness and fairness, revealing weaknesses in the model's effectiveness against adversarial attacks. Suggestions involve improving datasets and tackling dialect-related issues in NLP.

8.Offensive language detection in low resource languages: A use case of Persian language, Authors:Marzieh Mozafari, Khouloud Mnassri, Reza Farahbakhshi, and Noel Crespi

It offers a new collection of 6,000 annotated Twitter microblog posts to assist in this detection. A range of machine learning (ML), deep learning (DL), and transformer-based models were utilized, with an ensemble model demonstrating notable performance enhancements. The research highlights the significance of feature extraction methods, such as TF-IDF and word embeddings, in improving model precision. Future efforts will focus on tackling dataset imbalances and examining advanced language models to enhance the detection of offensive material.

9.Elevating Offensive Language Detection: CNN-GRU and BERT for Enhanced Hate Speech Identification, Authors:M.Madhavi, Dr. Sanjay Agal, Niyati Dhirubhai Odedra, Harish Chowdhary

The research article introduces an innovative method for identifying hate speech through the combination of Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), and BERT models. This combined architecture successfully captures local and sequential attributes, along with contextual representations, improving the detection of offensive material. The approach entails preparing a dataset of 24,783 tweets, then training the CNN-GRU and adjusting the BERT model. Experimental findings show comparable performance, reaching 98% accuracy in relation to current techniques, thereby enhancing a safer online environment.

10.ToxiCloakCN: Evaluating Robustness of Offensive Language Detection in Chinese with Cloaking Perturbations, Authors:Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, Roy Ka-wei Lee

The research article presents ToxiCloakCN, a dataset aimed at assessing the resilience of large language models (LLMs) in identifying offensive language in Chinese, especially in the context of homophonic and emoji variations. The research shows that current models, such as GPT-4o, greatly lag behind when confronted with these hidden malicious materials, underscoring a weakness in their ability to detect them. It highlights the necessity for sophisticated methods to enhance detection and more closely replicate human comprehension of subtle language nuances. Limitations comprise the dataset's failure to encompass all adversarial strategies and its dependence on the ToxiCN dataset, which might not adequately reflect linguistic variety.

III. PROBLEM STATEMENT

This research aims to create a text classification model for identifying offensive language on social media through enhanced preprocessing, embedding methods, and machine learning. Through the automation of content moderation, the model intends to improve detection precision and encourage safer online spaces.

The current system mentioned in the article emphasizes the issues of abusive language on social media channels, with a specific focus on cyberbullying and digital harassment. The growth of harmful content on social media, reflected by the rising instances of cyberbullying, endangers the mental and physical well-being of users. The document recognizes that multiple methods, employing deep learning and natural language processing strategies, have been created to tackle the identification of offensive language across diverse contexts and platforms. These methods frequently focus on particular types of harmful language, like hate speech, and are assessed with various datasets.

Drawbacks of Existing System :-

Platform Specificity: The model might not perform effectively across all social media platforms.

Challenges in Data Preprocessing: Managing slang and casual language may pose difficulties.

Generalization problems: It's difficult to apply the model effectively across various platforms

IV. MOTIVATION

To tackle the pervasive problem of harmful language on social media that impacts people and communities through harassment, hate speech, and emotional turmoil, automated moderation systems are crucial, particularly for large platforms like Twitter with extensive user bases. Current studies have investigated the automatic identification of abusive language; however, existing solutions tend to be limited in focus. This suggested system intends to enhance harmful content detection by combining multiple preprocessing approaches, embedding techniques, and classifiers, along with hyperparameter tuning. This all-encompassing method improves detection rates and provides a more efficient solution for managing online communities, thus minimizing the harmful effects of abusive content on users.

In response to the extensive use of offensive language on social media that impacts both individuals and communities, the proposed system aims to tackle bullying, hate speech, and other harmful content, which lead to significant emotional distress for users facing social exclusion and mental anguish. Given the vast number of users on Twitter and other platforms, this moderation system would be extremely difficult to handle without the use of automation.

Likewise, the research community has restricted itself to the same issue, investigating different methods in the field of automatic prevention of abusive language. Nonetheless, the solutions suggested to address this issue often appear restricted and focused in their approach regarding the type of offensive language. It challenges the creation of a better and more evolved solution to tackle the problems associated with social media text.

Considering that a model is supplied that will employ various preprocessing methods, diverse embedding techniques, and classification approaches, the system anticipates a higher detection rate and indicates a more efficient way of identifying harmful content. These offer a variety of machine learning algorithms along with their hyperparameter optimization features that allow the model to achieve remarkable detection abilities, thereby being applicable in effectively moderating online communities to reduce the effects caused by abusive language on individuals.

V. RESEARCH GAP

There is scant research on intricate assignments such as sequence labeling with unbalanced classes in NLP and Bio-Informatics. The article highlights issues in the portrayal of crime texts and proposes avenues for upcoming research. The necessity for specialized knowledge in data annotation is emphasized, revealing a lack of available resources.

The paper introduces OptSLA to tackle the issues in sequential labeling tasks, highlighting a deficiency in current aggregation techniques. The assessment of generative adversarial networks for representing crime text data indicates a deficiency in existing methods.

VI. OBJECTIVE

The developed system aims to establish a stronger model for the automated identification of offensive language on social media platforms, particularly Twitter, with the following objectives: improving precision to create a system that effectively detects offenses by utilizing various machine learning techniques in unison; sophisticated preprocessing, varied embedding methods; and multiple classification algorithms.

Enhance detection: Assess various embedding techniques, such as TF-IDF, along with classifiers like AdaBoost, SVM, and MLP, while fine-tuning the model to maximize performance in identifying offensive content, especially focusing on their hyperparameters to achieve the best accuracy.

Reduce Emotional Harm: To combat emotional harm, such as bullying or hate speech directed at individuals, an efficient tool for automatic content moderation on social media platforms should be made available.

The suggested task will enhance the abilities of social media moderators by enabling automated systems to manage the scale and fluctuations of social media content, thereby offering scalable approaches for identifying abusive language across various online settings.

Research Contribution: This study aims to advance current research in the area.

VI. METHODOLOGY

Step 1: Data Collection and Preparation

- Data Source:Gather a dataset of social media contributions, particularly from platforms such as Twitter, recognized for user-generated content that could feature offensive language.
- Data Labeling: Make sure the dataset is annotated, with each item classified as either offensive or non-offensive. If a labeled dataset is not accessible, think about employing human annotators or crowdsourcing to tag the data.
- Data Cleaning:

Remove Noise: Remove URLs, hashtags, mentions, and unnecessary special characters from the text to enhance its clarity for analysis.

Lowercasing: transform all text to lowercase for consistency.

- Tokenization: Divide the text into separate tokens (words) to ease embedding.
- Stopwords Removal: Eliminate frequent stopwords (e.g., "the", "and") lacking substantial meaning.
- Lemmatization/Stemming: Simplify words to their base or root form to standardize the vocabulary (e.g., "running" → "run").

Step 2: Feature Extraction (Embedding Techniques)

Embedding Techniques Overview: Apply various embedding techniques to transform text data into numerical features suitable for machine learning models.

Method 1: TF-IDF (Term Frequency-Inverse Document Frequency):

Compute the term frequency in every document while diminishing the importance of frequently used terms throughout the dataset to emphasize more significant words.

TF-IDF aids in illustrating the significance of words in relation to specific documents and the entire corpus.

Method 2: Word Embeddings (e.g., Word2Vec or GloVe):

Capture semantic connections among words by generating compact vector representations.

Method 3: Bag-of-Words (BoW):

Generate a vector representation of text based on the frequency of words while disregarding word order and meaning.

- Embedding Selection: TF-IDF was selected due to its simplicity and efficiency in text classification tasks, particularly when the context of words is not as important as their frequency.

Step 3: Model Selection and Implementation

- Choice of Classifiers: Utilize various classifiers to determine the most efficient model for identifying offensive language.
- Support Vector Machines (SVM): Performs effectively in high-dimensional settings and handles large feature sets efficiently.
- Multi-Layer Perceptron (MLP): A kind of neural network capable of learning intricate relationships.
- AdaBoost: A method that iteratively merges weak classifiers to enhance model effectiveness.
- Implementation Tools: Utilize Python libraries like Scikit-learn, TensorFlow, or Keras to create and train these models.

Step 4: Model Training and Evaluation

- Training: Utilize the chosen embeddings and fine-tuned hyperparameters to train each classifier.
- Testing: Assess the models using a distinct test set to measure their performance in real-world scenarios.
- Performance Comparison: Compare the outcomes of all classifiers to determine the optimal combination. The experiments conducted in this research revealed that the integration of AdaBoost, SVM, and MLP with the TF-IDF embedding yielded the highest average F1-scores.

Step 5: Results and Analysis

- Analysis of Results: Show the mean F1-scores for every model, and examine the confusion matrix to determine prevalent misclassification trends.
- Model Insights:
- The TF-IDF embedding technique effectively emphasized key terms in the dataset, leading to enhanced performance.
- Ensemble techniques such as AdaBoost, when paired with strong classifiers like SVM and MLP, improved the stability and precision of the models.
- Visualization: Utilize visuals (e.g., ROC curves, bar charts comparing F1-scores) to demonstrate the performance variances between classifiers.

Step 6: Model Deployment and Future Improvements

- Deployment Considerations:

Deploy the model in a live setting utilizing a scalable framework (e.g., Flask, FastAPI) for immediate detection.

- Future Enhancements:

Investigate deep learning methods like Transformer models (e.g., BERT) for possibly enhanced context comprehension.

Augment the dataset and incorporate additional social media platforms to enhance model generalization.

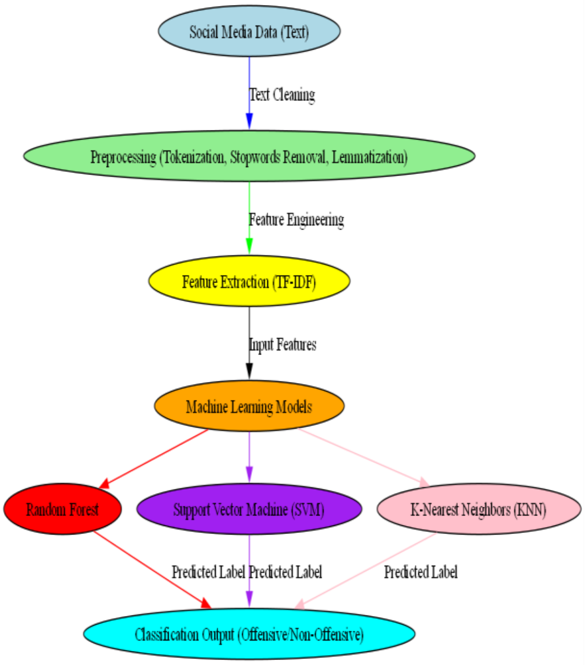
Social Media Data (Text)
Text Cleaning
Preprocessing (Tokenization, Stopwords Removal, Lemmatization)
Feature Engineering
Feature Extraction (TF-IDF)
Input Features
Machine Learning Models
Random Forest
Support Vector Machine (SVM)
K-Nearest Neighbors (KNN)
Predicted Label
Predicted Label
Predicted Label
Classification Output (Offensive/Non-Offensive)

Figure 1: Architecture

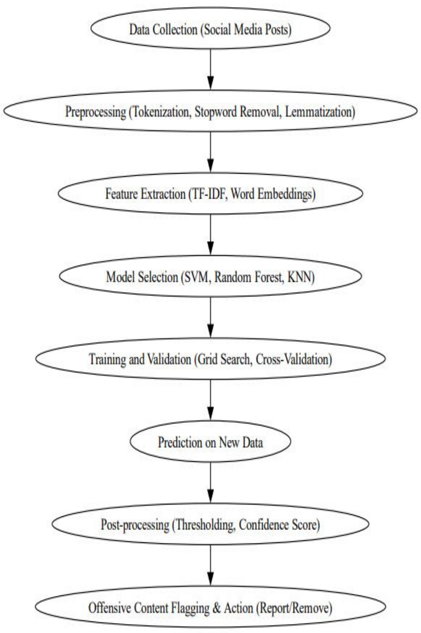
Data Collection (Social Media Posts)
Preprocessing (Tokenization, Stopword Removal, Lemmatization)
Feature Extraction (TF-IDF, Word Embeddings)
Model Selection (SVM, Random Forest, KNN)
Training and Validation (Grid Search, Cross-Validation)
Prediction on New Data
Post-processing (Thresholding, Confidence Score)
Offensive Content Flagging & Action (Report/Remove)

Figure 2: Proposed Flow

VII. Result

The evaluation metrics used to assess the performance of different classification algorithms—Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN)—are Accuracy, Precision, Recall, and F1-Score. Here's an explanation of each, along with their relevance and performance interpretation based on the provided data:

1. Accuracy

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

2. Precision

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

3. Recall

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

4.F1-Score

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

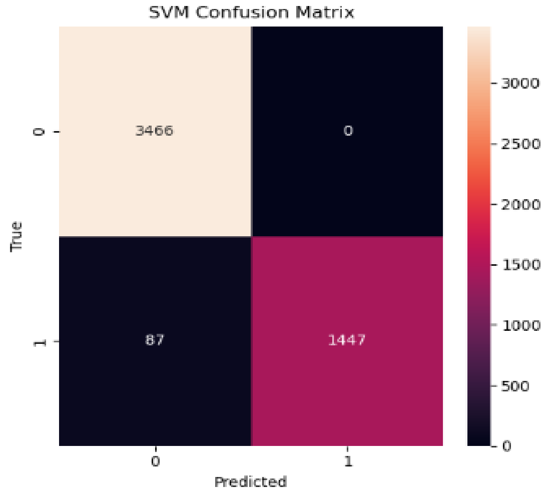

Figure 3: SVM Confusion Matrix

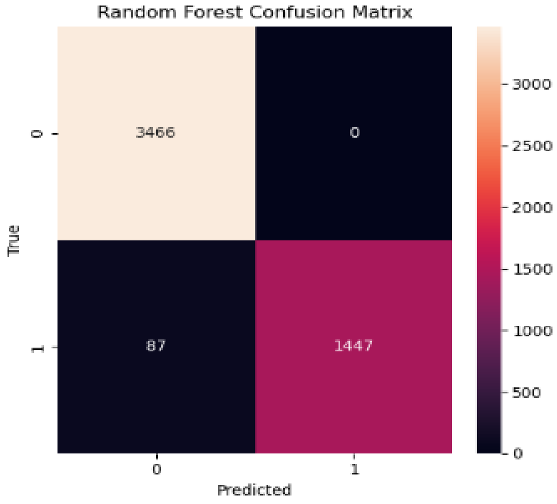

Figure 4: RF Confusion Matrix

Table 1: Classifier Result

	SVM	Random Forest	KNN
Accuracy	0.9826	0.9826	0.9826
F1-Score	0.9825	0.9825	0.9825

All three classifiers perform equally well in this experimental setup. Given such results, model selection may then depend on factors like computational efficiency, interpretability, or training time.

VIII. Conclusion

In this study, present a modular text classification pipeline aimed at social media datasets, specifically targeting Twitter. Our suggested method is to utilize a modular development that facilitates the straightforward integration of various text classification elements. The primary contribution of this paper is the introduction of a novel modular text classification pipeline aimed at enhancing benchmarking through a comprehensive analytical investigation of the top-performing techniques, features, and embeddings identified by the leading-edge research.

References

[1] Kogilavani Shanmugavadivel, V E Sathishkumar, Sandhiya Raja, T Bheema Lingaiah, S Neelakandan, and Malliga Subramanian, (2022), Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data.

[2] Anas Ahmed Raheeq and Mrs. Afroze Begum, (2024), OFFENSIVE LANGUAGE DETECTION

[3] Marcos Zampieri, S Rosenthal, Preslav Nakov, Alphaeus Dmonte, and Tharindu Ranasinghe, (2023), OffensEval 2023: Offensive language identification in the age of Large Language Models .

[4]Aiqi Jiang and A Zubiaga, (2023), Cross-lingual Offensive Language Detection: A Systematic Review of Datasets, Transfer Approaches and Challenges.

[5] Brillian Fieri and Derwin Suhartono, (2023), Offensive Language Detection Using Soft Voting Ensemble Model.

[6] Khouloud Mnassri, Reza Farahbakhsh, Razieh Chalehchaleh, (2024), A survey on multi-lingual offensive language detection.

[7] Zuchao Li, Min Peng, Israe Abdellaoui, Anass Ibrahimi, Mohamed Amine El Bouni, Asmaa Mourhir, Saad Driouech, and Mohamed Aghzal , (2024), Investigating Offensive Language Detection in a Low-Resource Setting with a Robustness Perspective.

[8] Marzieh Mozafari, Khouloud Mnassri, Reza Farahbakhshi, and Noel Crespi, (2024), Offensive language detection in low resource languages: A use case of Persian language.

[9] M.Madhavi, Dr. Sanjay Agal, Niyati Dhirubhai Odedra, Harish Chowdhary, (2024), Elevating Offensive Language Detection: CNN-GRU and BERT for Enhanced Hate Speech Identification.

[10] Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, Roy Ka-wei Lee, (2024), ToxiCloakCN: Evaluating Robustness of Offensive Language Detection in Chinese with Cloaking Perturbations.

INFORMATION ABOUT AUTHOR:

Dr.Konda Hari Krishna, Associate Professor, has more than 11 years of teaching experience in the department of computer science & engineering.
presently, he is working as an associate professor in the department of cse, school of computing, mohan babu university, tirupati, a.p. he received his Ph.D. in computer science & engineering from lingaya's vidyapeeth deemed to be university, faridabad, haryana. His research area is the improvement of network lifetime using clustering and dynamic topology methods in wsn. he is a good researcher & has published 14textbooks & 3 book chapters and worked mostly on wireless sensor networks, iot, ai & ml, ad hoc networks, network security, software engineering, mobile communications, dbms, data warehousing, data mining, big data & analytics, and cloud computing. He published various patents (16)& 35 research papers in various international journals of reputed & presented research papers in conferences & seminars, and editorial board member & reviewer for various journals & conferences.
he is an active member in academics & administrative activities &publication member in skrgc journal & igi global chapter & life member in technical bodies like ie, iste, csi, iaeng, ired, insticc. He has completed 11 nptel, 1 arpit & 3 cdac certifications on various latest technologies.he was recently awardedbest young faculty, researcher & academic & research excellence in 2022, 2023 & 2024.
E-mail:khk396@gmail.com,konda.harikrishna@mbu.asia

Information about author:

Ms. Vankadari Yasaswini is currently pursuing a master of computer applications (mca) at mohan babu university. he completed his bachelor of computer science (bsc) at sdhr degree and pg college. She has completed internships on web development. She is an active class representative in academics and team lead in books club coordination.

E-mail:vankadariyasaswini07@gmail.com