Enhanced Dark Web Classification Using LDA-Text CNN Feature Fusion and CNN2D

J.Kumari [1] , B.Gayathri [2],

[1] Asst.Professor, [2] PG.Scholar

Department of MCA, QIS College of Engineering & Technology, (Autonomous) Ongole, AP, India.

Abstract: Because it provides anonymity, the Dark Web serves as a platform for both legitimate and illicit activity. This work introduces a deep learning-based classification technique that uses TextCNN and Topic Modeling weights to enhance the detection of dangerous Dark Web services. Unlike traditional methods like TF-IDF and Latent Semantic Analysis, which cannot exclude unnecessary data, our approach employs Latent Dirichlet Allocation (LDA) to identify the most crucial topic features. The model is trained using a Kaggle dataset, and preprocessing is carried out to reduce noise and enhance data representation. When LDA-generated topic weights are input into a TextCNN model, classification accuracy is higher than when KNN and Random Forest are employed. An improved CNN2D model with Dropout layers is used to substantially improve classification performance by successfully removing superfluous features. Experimental results show that our approach is highly effective at identifying Dark Web services since it makes use of 90 perfect themes, which significantly increases accuracy.

Index terms *- Dark Web Classification, TextCNN, Topic Modeling, Latent Dirichlet Allocation (LDA), Deep Learning, TF-IDF, CNN2D, Dropout Layer, Dark Web Services Detection, Cybersecurity.*

1. INTRODUCTION

The Dark Web serves as a center for both legal and illegal activity due to its secrecy. Despite its usage for private conversation and privacy protection, cybercrimes such illegal trading, hacking, and data breaches also take use of it. Because it makes it simpler to monitor and stop unwanted activity, identifying and categorizing Dark Web services is essential for cybersecurity. However, because of its intricate language patterns and irrelevant data, typical machine learning techniques are unable to classify Dark Web material effectively.

To solve these issues, this paper suggests a DL classification model based on TextCNN and Topic Modeling weights. Our methodology uses Latent Dirichlet Allocation (LDA) to give often occurring words greater weights, in contrast to traditional methods like TF-IDF, Document Matrix, and Latent Semantic Analysis, which are unable to efficiently remove superfluous phrases. By integrating TextCNN with topic weights derived from LDA, the suggested method increases classification accuracy.

Data collection, preprocessing, feature extraction using LDA, and classification using TextCNN are only a few of the several stages involved in this work. For training and assessment, a Kaggle dataset including more than 10,000 Dark Web services is utilized. According to experimental findings, choosing the top 90 topic traits greatly enhances classification performance. To improve feature selection and boost overall accuracy, an extended CNN2D model is supplemented with a Dropout layer. The effectiveness of the proposed technique in accurately identifying Dark Web services is demonstrated by comparing it with other models, such as KNN and Random Forest.

2. LITERATURE SURVEY

a) A novel approach for dimension reduction using word embedding: An enhanced text classification approach:

Reducing the dimensional feature space is one of the difficult issues in text categorization. In order to identify words with comparable semantic meaning, this study presents an improved text classification method that uses the Bag-of-Words representation model with term frequency-inverse document frequency (tf-idf) and the word embedding technique "GloVe." The most representative term with comparable meanings is the one with the biggest total of tf-idf. Other techniques including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Latent Semantic Indexing (LSI), and a hybrid approach PCA+LDA employing the Naïve Bayes classifier are compared to the performance of the suggested method. The suggested approach outperforms current dimension reduction methods in classification, according to experimental findings on three datasets: the BBC, Classic4, and 20-newsgroup datasets. Finally, in order to assess the classifier's performance on the decreased features, we developed a new performance assessment metric..

b) Multi-label Text Classification Using Semantic Features and Dimensionality Reduction with Autoencoders:

In order to lower the large dimensionality of feature space in text classification, feature selection is crucial. The many statistical methods that have been put out for feature selection and weighting suffer from the loss of semantic relationships between ideas as well as the disregard for dependencies and word order. Two methods for integrating semantics into feature selection are presented in this paper. Additionally, we analyze the performance penalty of feature extraction by transforming the features into a smaller feature space using autoencoders. The Bag-of-Word (BOW) frequency based feature selection approach utilizing term frequency/inverse document frequency (TF-IDF) for feature weighting is much outperformed by semantic-based feature selection strategies, according to our extensive studies using the EUR-lex dataset. Additionally, the autoencoders are still able to provide superior features than BOW with TF-IDF even after an extensive dimensionality reduction of the original features with a factor of 10.

c) Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM)

The proliferation of false news, whether produced by humans or algorithms, has a detrimental impact on society and individuals on both a political and social level. The rapid cycling of news in the age of social media makes it difficult to quickly assess its dependability. As a result, automated methods for

identifying bogus news have become essential. Principle Component Analysis (PCA) and Chi-Square are two distinct dimensionality reduction techniques used with a hybrid neural network architecture that combines the capabilities of CNN and LSTM to handle the aforementioned problem. Before sending the feature vectors to the classifier, this paper suggested using dimensionality reduction techniques to lower their dimensionality. This study used a dataset from the Fake News Challenges (FNC) website, which has four different attitude types: agree, disagree, debate, and irrelevant, to create the rationale. For the purpose of detecting bogus news, the nonlinear characteristics are input into PCA and chi-square. Finding out how a news story feels about its headline is the driving force behind this study. The accuracy and F1-score values are improved by approximately 4% and 20%, respectively, using the suggested model. With 97.8% accuracy, the testing findings demonstrate that PCA performs better than Chi-square and cutting-edge techniques.

d) Threats from the Dark: A Review over Dark Web Investigation Research for Cyber Threat Intelligence:

Analyzing Dark Web information is crucial to preventing cybercrimes and comprehending criminal thoughts, from preemptive cyberattack detection to identifying important individuals. Whether doing a stand-alone investigation of the Dark Web or an integrated one that incorporates information from the Surface Web and the Deep Web, research in the Dark Web has shown to be a crucial step in combating cybercrime. In this study, we examine recent research on the analysis of Dark Web content for Cyber Threat Intelligence (CTI), offering a thorough examination of their methodology, tools, approaches, and outcomes as well as a discussion of any potential drawbacks. In this review, we show how important it is to examine the material of various Dark Web platforms, guiding new researchers via cutting-edge techniques. We also go into the domain's future directions, ethical issues, and technological difficulties.

e) Phishing Web Page Detection with HTML-Level Graph Neural Network

One of the biggest risks to Internet users is phishing websites. Conventional techniques for detecting phishing websites depend on traits that are created by hand. Deep learning-based techniques that use HTML as input have recently significantly improved detection performance. They often use Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) for classification, using HTML codes as character sequences. Nevertheless, CNN and RNN usually fail to simulate the long-range semantics that is essential for phishing detection and can only extract local information from the HTML code sequences. In this study, we offer a novel phishing web site detection algorithm based on Graph Neural Networks (GNNs) that can efficiently use HTML's intrinsic structural information to capture long-range semantics. First, we use the Document Object Model (DOM) to naturally represent an HTML as a graph, and then we use RNN to extract the local characteristics of node properties. Next, using these local properties and the graph structure, we use GNN to simulate the long-range relationships between nodes. To better comprehend the intent of HTML codes, our suggested model combines the benefits of RNN and GNN. Our method's accuracy significantly exceeds existing state-of-the-art

approaches, as shown by extensive testing on a real-world dataset.

3. METHODOLOGY

i) Proposed Work:

This paper provides an extended deep learning model based on the top-performing LDA-TextCNN approach to increase the classification accuracy of Dark Web services. The suggested approach trains a TextCNN model for classification after obtaining topic modeling features using Latent Dirichlet Allocation (LDA). We augment this model by using CNN2D, which improves classification accuracy and fine-tunes feature selection, to further boost performance.

This expanded method gives the CNN2D network the best characteristics from the LDA-TextCNN model. CNN2D enhances the model's capacity to distinguish between pertinent and irrelevant information by employing convolutional layers to better capture spatial correlations among topic features. In order to ensure that only the most important topic weights contribute to classification, a Dropout layer is also included to eliminate noisy and duplicate features.

When compared to current techniques like KNN and Random Forest, the extended model successfully lowers misclassification rates and achieves improved accuracy by integrating LDA-TextCNN with CNN2D and Dropout. According to experimental findings, this method greatly improves Dark Web service detection, which makes it a good option for cybersecurity applications.

ii) System Architecture:

This paper presents an expanded deep learning model based on the top-performing LDA-TextCNN approach to increase the classification accuracy of Dark Web services. The suggested technique trains a TextCNN model for classification after obtaining topic modeling features using Latent Dirichlet Allocation (LDA). We augment this model by using CNN2D, which improves classification accuracy and optimizes feature selection, in order to further boost performance.

The CNN2D network receives the best characteristics from the LDA-TextCNN model in this expanded method. CNN2D enhances the model's capacity to distinguish between relevant and irrelevant information by employing convolutional layers to better capture spatial correlations among topic features. To ensure that only the most important topic weights contribute to classification, a Dropout layer is also included to eliminate noisy and duplicate features.

Compared to current techniques like KNN and Random Forest, the extended model successfully lowers misclassification rates and achieves greater accuracy by integrating LDA-TextCNN with CNN2D and Dropout. According to experimental findings, this method greatly improves Dark Web service detection, making it a good option for cybersecurity applications.

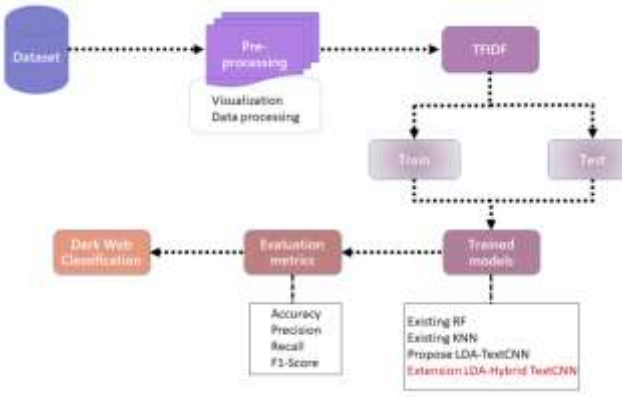

Fig.1. Proposed Architecture

iii) MODULES:

i. *Data Loading:*
   - Imports the dataset for training and testing.
   - Uses a Kaggle dataset containing over 10,000 Dark Web service descriptions.

ii. *Visualization:*
   - Generates graphs to display different Dark Web services.
   - X-axis represents service names, and Y-axis indicates their counts.

iii. *Data Preprocessing:*
   - Removes digits, special symbols, and stop words.
   - Applies stemming and lemmatization to standardize words.

iv. *TF-IDF Transformation:*
   - Converts processed text into numerical vectors.
   - Assigns word frequency scores for better feature extraction.

v. *LDA Topic Modeling:*
   - Extracts the most relevant words from each sentence.
   - Selects the top 90 topic weights for classification.

vi. *Data Splitting:*
   - Divides the dataset into training and testing sets.
   - Ensures an optimal split for model evaluation.

vii. *Model Generation:*
   - *Existing Models***:** KNN and Random Forest for baseline comparison.
   - *Proposed LDA-TextCNN:* Uses LDA topic weights with TextCNN for classification.
   - *Extension LDA-Hybrid TextCNN:*
     - Extracts best features from LDA-TextCNN.
     - Trains a CNN2D model on extracted features.
     - Uses a Dropout layer to remove irrelevant features.
     - Enhances classification accuracy.

viii. *Admin Login:*
   - Allows admin authentication and access to system functionalities.

ix. *Dark Web Classification:*
   - Enables users to upload test data.
   - Classifies Dark Web services based on the trained models.

x. *Prediction:*
   - Displays final classification results.
   - Shows accuracy improvements using the extension model.

iv) ALGORITHMS:

a) KNN

Dark Web services are categorized by the K-Nearest Neighbors (KNN) algorithm according to how closely their characteristics match labeled training data points. KNN forecasts a service's class label using feature space distance measurements. In order to evaluate the accuracy and efficacy of KNN with more sophisticated algorithms, this study employs it as a baseline classification model.

This aids in evaluating the overall performance gains brought about by the application of LDA-TextCNN and other suggested models.

b) Random Forest:

During training, several decision trees are constructed using the Random Forest ensemble learning technique, which then outputs the mode of the trees' predictions for classification tasks. It is suitable for the Dark Web categorization effort because it reduces overfitting and effectively manages high-dimensional data. The algorithm uses Random Forest to identify intricate patterns in the data, resulting in accurate predictions for the categories of Dark Web services. The performance of this algorithm is used as a standard for assessing suggested approaches.

c) LDA-TextCNN:

To categorize Dark Web services, the suggested LDA-TextCNN model combines a Convolutional Neural Network (CNN) with Latent Dirichlet Allocation (LDA) topic modeling. LDA improves feature representation by extracting topic weights from the processed text. The TextCNN then uses these enhanced characteristics to do classification, increasing its robustness and accuracy.

When compared to more conventional techniques like KNN and Random Forest, this strategy seeks to better capture the semantic relationships present in the data, which will eventually produce superior classification results.

d) Extension LDA-Hybrid Texting:

By adding a second CNN2D layer that was trained using the best features from the original model, the extension LDA-Hybrid TextCNN model expands upon the suggested LDA-TextCNN. By improving feature representation and lessening the influence of irrelevant data, this hybrid technique seeks to improve classification accuracy.

This model's use of dropout layers reduces overfitting and produces a more accurate categorization of Dark Web services. The goal of this improved model is to outperform previous algorithms.

4. EXPERIMENTAL RESULTS

A Kaggle dataset with over 10,000 descriptions of Dark Web services was used to assess the suggested solution. To enhance text quality, data preparation methods including lemmatization, stemming, and stop word removal were used. Latent Dirichlet Allocation (LDA) was utilized to extract significant topic features after TF-IDF was employed for the initial feature representation. The best results were obtained by experimenting and choosing the top 90 subjects. To achieve objective evaluation, the dataset was divided into training and testing sets. Several models, such as KNN, Random Forest, LDA-TextCNN, and the extended LDA-Hybrid TextCNN (CNN2D with Dropout), were used for comparison.

The findings show that the suggested models perform noticeably better than conventional methods. The LDA-TextCNN model increased accuracy to 95.05%, while KNN and Random Forest attained accuracies of 83.70% and 87.46%, respectively. The maximum accuracy of 96.09% was attained by further improvement utilizing the LDA-Hybrid TextCNN (CNN2D with Dropout), coupled with better precision, recall, and F1-score. Better generalization resulted from the successful reduction of noise and overfitting caused by the combination of CNN2D and Dropout. These results demonstrate that the suggested hybrid deep learning model offers a very effective and dependable method for classifying Dark Web material..

Accuracy: A test's accuracy is determined by its capacity to distinguish between instances of health and illness. To gauge the accuracy of the test, find the percentage of analyzed instances that had true positives and true negatives. According to the computations:

Accuracy = TP + TN /(TP + TN + FP + FN)

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

Test Accuracy: 0.9895

Precision: Precision is the number of positive cases or the accuracy rate of a categorization. The following formula is used to calculate accuracy:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\ cision = \frac{TP}{(TP + FP)}$$

Recall: A model's ability to recognize every instance of a pertinent machine learning class is measured by its recall. The ratio of accurately predicted positive observations to the total number of positives indicates how well a model detects class instances.

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

mAP: It considers the number of relevant recommendations and their position on the list. MAP at K is calculated using the arithmetic mean of the Average Precision (AP) at K for each user or query.

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

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

$$n = \text{the number of classes}$$

F1-Score: An accurate machine learning model is indicated by a high F1 score. combining precision and recall to increase model correctness. The accuracy statistic indicates how frequently a model correctly predicts a dataset.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\ cision)}{(Recall + \Pr e\ cision)}$$

Algorithm Name	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)
Existing KNN	83.7	82.76	82.92	82.75
Existing Random Forest	87.46	86.74	86.36	86.41
Proposed LDA-TextCNN	95.05	94.93	94.19	94.34
Extension LDA-Hybrid TextCNN	96.09	95.83	95.94	95.55

Fig.7. Comparison table of performance evaluation metrics of all algorithms

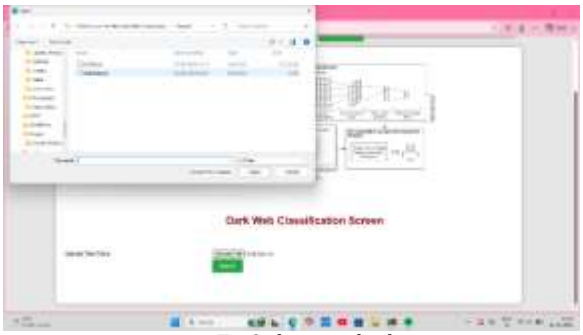

Fig.8. dataset upload page

Fig.9. classification service page

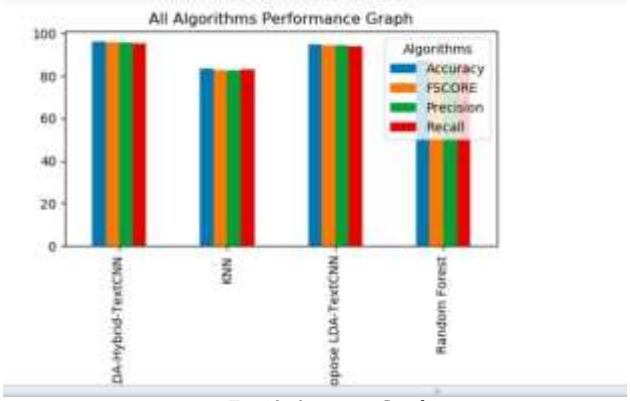

Fig.10. Accuracy Graph

5. CONCLUSION

This study used an enhanced CNN2D model and LDA-weighted TextCNN to offer an effective deep learning-based method for Dark Web content categorization. Overcoming the drawbacks of conventional techniques like TF-IDF and LSA, the combination of Latent Dirichlet Allocation (LDA) with TextCNN enhanced feature representation by collecting significant topic-based information. Better feature refinement and less overfitting were made possible by further improvement utilizing CNN2D with a Dropout layer, which enhanced classification performance.

According to experimental findings, the suggested hybrid model performs noticeably better in terms of accuracy, precision, recall, and F1-score than traditional machine learning algorithms like KNN and Random Forest. The technology is quite successful in identifying and categorizing Dark Web services, with an accuracy of 96.09%. All things considered, the suggested method offers cybersecurity applications a reliable and scalable solution that makes it possible to effectively monitor and identify potentially harmful activity on the Dark Web.

6. FUTURE SCOPE

By adding sophisticated deep learning models like Transformers and attention-based architectures (like BERT) to extract deeper contextual associations from Dark Web text data, the suggested system can be further improved. Furthermore, using multilingual text analysis and real-time data collecting from active Dark Web sources can enhance the system's adaptability and resilience in a variety of settings.

By integrating the model into a real-time cybersecurity monitoring system, future research may also concentrate on enhancing scalability and deployment. Additionally, integrating text-based analysis with other modalities, including network traffic data or image-based material, can result in a more thorough Dark Web detection framework, allowing for the more precise identification of intricate and dynamic cyberthreats.

REFERENCES

[1] C. A. S. Murty and P. H. Rughani, ‘‘Dark Web text classification by learning through SVM optimization,’’ J. Adv. Inf. Technol., vol. 13, no. 6, 2022, doi: 10.12720/jait.13.6.624-631.

[2] A. H. M. Alaidi, R. M. Al Airaji, H. T. S. Alrikabi, I. A. Aljazaery, and S. H.

Abbood, ‘‘Dark Web illegal activities crawling and classifying using data mining techniques,’’ Int. J. Interact. Mobile Technol., vol. 16, no. 10, pp. 122–139, May 2022, doi: 10.3991/ijim.v16i10.30209.

[3] N. Deguara, J. Arshad, A. Paracha, and M. A. Azad, ‘‘Threat miner—A text analysis engine for threat identification using dark Web data,’’ in Proc. IEEE Int. Conf. Big Data, Dec. 2022, pp. 3043–3052, doi: 10.1109/Big Data55660.2022.10020397.

[4] Y. Jin, E. Jang, Y. Lee, S. Shin, and J.-W. Chung, ‘‘Shedding new light on the language of the dark Web,’’ 2022, arXiv:2204.06885.

[45] H. Ma, J. Cao, B. Mi, D. Huang, Y. Liu, and Z. Zhang, ‘‘Dark Web traffic detection method based on deep learning,’’ in Proc. IEEE 10th Data Driven Control Learn. Syst. Conf. (DDCLS), May 2021, pp. 842–847, doi: 10.1109/DDCLS52934.2021.9455619.

[6] K. N. Singh, S. D. Devi, H. M. Devi, and A. K. Mahanta, ‘‘A novel approach for dimension reduction using word embedding: An enhanced text classification approach,’’ Int. J. Inf. Manage. Data Insights, vol. 2, no. 1, Apr. 2022, Art. no. 100061, doi: 10.1016/j.jjimei.2022.100061.

[7] W. Alkhatib, C. Rensing, and J. Silberbauer, ‘‘Multi-label text classification using semantic features and dimensionality reduction with autoencoders,’’ in Proc. 1st Int. Conf. Lang., Data, Knowl., 2017, pp. 380–394, doi: 10.1007/978-3-319-59888-8_32.

[8] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, ‘‘Fake news stance detection using deep learning architecture (CNN-LSTM),’’ IEEE Access, vol. 8, pp. 156695–156706, 2020, doi: 10.1109/ACCESS.2020.3019735.

[9] R. Basheer and B. Alkhatib, ‘‘Threats from the dark: A review over dark Web investigation research for cyber threat intelligence,’’ J. Comput. Netw. Commun., vol. 2021, pp. 1–21, Dec. 2021, doi: 10.1155/2021/1302999.

[10] A. Alharbi, M. Faizan, W. Alosaimi, H. Alyami, A. Agrawal, R. Kumar, and R. A. Khan, ‘‘Exploring the topological properties of the tor dark Web,’’ IEEE Access, vol. 9, pp. 21746–21758, 2021, doi: 10.1109/ACCESS.2021.3055532.

[11] L. Ouyang and Y. Zhang, ‘‘Phishing Web page detection with HTML level graph neural network,’’ in Proc. IEEE 20th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom), Oct. 2021, pp. 952–958, doi: 10.1109/TrustCom53373.2021.00133.

[12] H. Alnabulsi and R. Islam, ‘‘Identification of illegal forum activities inside the dark net,’’ in Proc. Int. Conf. Mach. Learn. Data Eng. (iCMLDE), Dec. 2018, pp. 22–29.

[13] D. Hayes, F. Cappa, and J. Cardon, ‘‘A framework for more effective dark Web marketplace investigations,’’ Information, vol. 9, no. 8, p. 186, Jul. 2018, doi: 10.3390/info9080186.

[14] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and L. Fernández-Robles, ‘‘ToRank:

Identifying the most influential suspicious domains in the tor network,'' Expert Syst. Appl., vol. 123, pp. 212–226, Jun. 2019, doi: 10.1016/j.eswa.2019.01.029.

[15] M. W. Al Nabki, E. Fidalgo, E. Alegre, and D. Chaves, ''Supervised ranking approach to identify infLuential websites in the darknet,'' Int. J. Speech Technol., vol. 53, no. 19, pp. 22952–22968, Oct. 2023, doi: 10.1007/s10489-023-04671-9.

Author profiles

Mrs. Jasti kumari is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. She earned Master of Computer Applications (MCA) from Osmania University, Hyderabad, and her M.Tech in Computer Science and Engineering (CSE) from Jawaharlal Nehru Technological University, Kakinada (JNTUK). Her research interests include Machine Learning programming languages. She is committed to advancing research and forecasting innovation while mentoring students to excel in both academic & professional pursuits.

B.Gayathri is an MCA Student in the Department of Computer Application at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. She has Completed Degree in B.Sc.(computers) from G.C.&Y.P.N Degree College Kanigiri, Prakasam district. Her area of interest are DBMS, Data analyst.