STM JOURNALS

Journal of
Mobile Computing, Communications & Mobile Networks

ISSN: 2349-901X
Volume 11, Issue 2, 2024
May–August
DOI (Journal): 10.37591/JoMCCMN

https://journals.stmjournals.com/jomccmn

Review JoMCCMN

# A Supervised Learning Approach for Toxic Comment Detection on Social Media Platforms

M. Prasad[1,*], Rajarao PBV[2], P. Kiran Sree[3], P. Gowthami[4], K. Ajita Lakshmi[5], B. Subrahmanyam[6]

***Abstract***

*Nowadays everyone uses social media platforms like X (formerly Twitter), Instagram, Facebook, etc. for various purposes. With the help of this, we share our opinions, ideas, and feelings. Generally, the datasets obtained from the internet are constructive; however, there is a significant proportion of toxic ones. The datasets are filtered to remove noise, and noise is removed in post-processing. The study initiates with the upload and preprocessing of a toxic comment dataset meticulously cleaning text by eliminating stop words and special symbols to lay out a standardized corpus. The resulting application of count vectorizer captures word occurrences constructing a feature matrix for algorithmic training supervised algorithms, including support vector machine, logistic regression, naive Bayes, random forest, decision tree, and k-nearest neighbors. These are systematically implemented and each algorithm undergoes rigorous assessment with accuracy measurements computed to check its proficiency in segregating toxic from non-toxic comments. The examination finishes in an accuracy graph visually contrasting the performance of the various supervised algorithms. This visual representation helps in identifying the most effective model for online toxic comment classification. The multiheaded model consists of toxicity, severe-toxic, obscene threat insults, and toxicity prediction based on confusion metrics. The practical implications of this study lie in outfitting a robust tool for online platforms to automatically detect and manage toxic comments adding to a safer and more constructive digital environment.*

***Author for Correspondence**
M. Prasad
E-mail: prasads.maddula@gmail.com

[1,2]Associate Professor, Department of Computer Science Engineering, Shri Vishnu Engineering College for Women (A), Bhimavaram, Andhra Pradesh, India
[3]Professor and Head, Department of Computer Science Engineering, Shri Vishnu Engineering College for Women (A), Bhimavaram, Andhra Pradesh, India
[4]Student, Department of Computer Science Engineering, Shri Vishnu Engineering College for Women (A), Bhimavaram, Andhra Pradesh, India
[5]Assistant Professor, Department of Electronics and Communication Engineering, Shri Vishnu Engineering College for Women(A), Bhimavaram, Andhra Pradesh, India
[6]Assitant Professor, Department of MCA, BVC Institute of Technology and Science(A), Batlapalem, Andhra Pradesh, India

Received Date: May 03, 2024
Accepted Date: May 09, 2024
Published Date: July 05, 2024

**Citation:** M. Prasad, Rajarao PBV, P. Kiran Sree, P. Gowthami, K. Ajita Lakshmi, B. Subrahmanyam. A Supervised Learning Approach for Toxic Comment Detection on Social Media Platforms. Journal of Mobile Computing, Communications & Mobile Networks. 2024; 11(2): 7–14p.

**Keywords:** Toxic comments, support vector machine (SVM), k-nearest neighbors (KNN), multiheaded model, supervised learning

## INTRODUCTION

The rapid advancement of computer science and technology has revolutionized 21st-century internet, enabling global communication through smartphones and the internet. Initially, email communication was filled with spam, making it difficult to classify emails as positive or negative. In the vast digital landscape, user-generated content has become a significant component of communication platforms. However, with this increased interaction comes the issue of toxic comments – those that include offensive language, hate speech, or other forms of harmful content. Since social media sites have substantially altered communication and data flow, it is becoming increasingly important to categorize content in a positive and negative way in

 

order to control antisocial behavior and prevent harm to society. Recent incidents of authorities arresting people for harmful social media content have highlighted the need to detect such content before it is published. Toxic comment classification using machine learning involves the application of algorithms to automatically identify and categorize comments that exhibit harmful or offensive behavior.

This process aims to create a safer and more inclusive online environment by filtering out content that may be abusive, discriminatory, or inappropriate. Machine learning models are trained on labeled datasets, where comments are categorized as toxic or non-toxic based on predefined criteria. To classify toxic comments on social media, different machine learning algorithms will be used, including logistic regression, random forest, support vector machine (SVM) classifier, naive Bayes, decision tree, and K-nearest neighbors (KNN) classification. These algorithms will be applied to the Kaggle.com data set and compared for accuracy, log loss, and hamming loss. Preprocessing steps, such as tokenization, stemming, and removing stop words, are applied to streamline the text data before feeding it into the machine learning model.

## LITERATURE REVIEW

Social media platforms have become an integral part of contemporary society, facilitating global communication, expression, and interaction. However, within the vast spectrum of opinions and ideas shared on these platforms, a notable portion is marred by toxic content. Detecting and managing toxic comments has thus emerged as a critical endeavor to foster a safer and more constructive digital space. Recent studies have explored diverse methodologies to tackle the challenge of toxic comment detection, leveraging advancements in machine learning and natural language processing.

For instance, Badjatiya et al. [1] delved into deep learning techniques for hate speech detection in tweets, demonstrating the effectiveness of neural network models in discerning toxic language patterns. This highlights the potential of sophisticated algorithms in capturing subtle linguistic cues indicative of toxicity. A common obstacle in toxic comment detection is class imbalance, where the number of toxic comments is significantly lower than non-toxic ones.

Prusa et al. [2] proposed random undersampling to address this issue in sentiment analysis of tweets, enhancing the robustness of machine learning models in distinguishing toxic content from benign comments by balancing class distribution.

Waseem et al. [3] introduced a comprehensive typology of abusive language detection subtasks, offering a systematic framework to categorize and understand abusive behaviors online. This approach facilitates nuanced detection strategies tailored to specific subtypes of abuse, enriching the arsenal of tools for combating toxic comments.

Furthermore, Rustam et al. [4] conducted a comparative analysis of supervised machine learning models for sentiment analysis of COVID-19 tweets, emphasizing the significance of algorithm selection in optimizing classification accuracy and robustness amidst evolving global contexts. To enhance the effectiveness of toxic comment detection models, addressing class imbalance and feature sparsity is imperative.

Fatima et al. [5] proposed a methodology to minimize overlapping degree and improve class-imbalanced learning under sparse feature selection, bolstering the discriminative power of machine learning models in identifying toxic comments amid noisy data.

Moreover, Rustam et al. [6] explored sentiment-based classification of tweets for US airline companies, showcasing the versatility of sentiment analysis techniques in understanding customer perceptions across different domains, including toxic comment detection on social media platforms. Practical text analytics methodologies offer valuable insights into the extraction and interpretation of textual data from social media platforms.

 

Anandarajan et al. [7] provided a comprehensive overview of such techniques, emphasizing the importance of maximizing the value of text data through advanced analytics and data science methodologies to develop robust frameworks for toxic comment detection and management.

In summary, recent advancements in machine learning, natural language processing, and text analytics present promising avenues for addressing the challenge of toxic comment detection on social media platforms. By harnessing sophisticated algorithms, mitigating class imbalance, and adopting comprehensive detection strategies, researchers strive to cultivate a safer and more conducive digital environment for users worldwide.

## PROPOSED METHODOLOGY

The field of machine learning is concerned with the development of computer programs that automatically improve with experience. It is a subset of artificial intelligence that focuses on optimizing performance based on past data or experience [8]. The data that is given for training the model is processed by the machine learning algorithms that are suited for the model. Generally vast amount of data is used for training the machine learning models. Used supervised machine learning algorithms for this toxic comment classification, such as SVM, logistic regression, decision trees, KNN, and naive Bayes.

- *Logistic regression:* Used for binary classification problems, predicting whether an instance belongs to one of two classes.
- *Support vector machine (SVM):* Effective for both classification and regression tasks. SVM aims to find a straight line that classifies the data into different categories.
- *Decision trees:* Tree-like models where all the sub-nodes represent a decision based on some parameters which lead to final output at the end.
- *K-nearest neighbor (KNN):* Classifies a new instance based on the highest instances of its K nearest neighbors in the variable space.
- *Naive Bayes:* A probabilistic algorithm based on Bayes' theorem, is commonly used for classification problems. It assumes independence between features.
- *Random forest:* A collective learning technique that constructs several decision trees and aggregates their forecasts to enhance precision and resilience.

The existing structure utilizes naive Bayes in text classification. Naive Bayes assigns posterior probabilities based on the occurrence of various words in texts, making computations more efficient than non-naive Bayes approaches. It is noted that naive Bayes is less accurate for text categorization, and one limitation is the assumption of independent features. In reality, it is unlikely that all attributes are completely independent. Despite not using the term frequency inverse document frequency (TF-IDF) idea for classification, the proposed method combines TF-IDF, content rank approach to enhance text categorization [9].

When it comes to toxic comment classification, the naive Bayes model, particularly the multinomial naive Bayes variant, is effective in categorizing harmful comments. The process involves preprocessing a comments dataset, tokenizing them, and removing common words that do not contribute significantly to classification. The naive Bayes model relies on Bayes' theorem and the assumption of independence between words in the comments, simplifying the prediction process. The model creates a probabilistic model to determine the likelihood of a harmful comment based on its wording.

Feature engineering is crucial, considering word frequency and relevance to toxicity. Laplace smoothing is often used to address the issue of zero probabilities for words not in a particular class during training. Once trained, the naive Bayes model is tested on different datasets using metrics like precision, recall, and F1 score to assess its ability to correctly classify toxic comments while minimizing false positives or negatives. Naive Bayes proves suitable for real-time toxic comment categorization, enabling quick moderation in content management. However, regular updates and adjustments are necessary to adapt the model to evolving emerging patterns of toxic behavior [10].

 

TF-IDF signifies term frequency inverse document frequency. It may be represented as the computation of how accurate a discussion is in the order or the compilation search out of the manual [11, 12]. The importance expands compared to the number of instances in the document a discussion is carried out but it is balanced by one–word repetition in the body of text (basic document file).

$$tfidf_{i,j} = tf_{i,j} * \log(N/df_i) \quad (1)$$

where
$tf_{i,j}$ = total number of occurrences of i in j
N = total number of documents
$df_i$ = total number of documents containing i

## RESULTS AND DISCUSSION

This paper initiates by addressing the crucial step of uploading the Toxic Comments Dataset, followed by a meticulous preprocessing phase to ensure data cleanliness. Each comment undergoes thorough examination, wherein systematic removal of stop words and special symbols is conducted. The frequency of occurrence for each word is then computed, leading to the construction of a count vector. This vector is subsequently employed for training with diverse algorithms, notably the SVM algorithm. To evaluate the effectiveness of these algorithms, the dataset is divided into an 80% subset for machine learning training and a 20% subset for testing machine learning performance. Furthermore, additional algorithms including logistic regression, naive Bayes, KNN, decision tree, and random forest are applied. A comprehensive analysis ensues, illustrating the accuracy and loss comparison across all algorithms. Finally, the test data is introduced, and the machine learning system predicts whether the data contains toxic comments. Figure 1 showcases the user interface, prompting the selection of the 'Upload Toxic Comments Dataset' button for dataset upload. Figure 2 depicts the process of selecting and uploading the 'train.csv' file, followed by loading the dataset by clicking the 'Open' button. Progressing further, Figure 3 demonstrates the phase where each comment is scrutinized for stop words and special symbols to ensure cleanliness. Subsequently, Figure 4 illustrates the selection and upload process of the 'testComment.csv' file, with the subsequent loading of the dataset by clicking the 'Open' button. Finally, Figure 5 exhibits the interface displaying the predicted result, indicating whether comments are classified as toxic or non-toxic.

An array of machine learning models is utilized for the classification of comments into toxic and non-toxic categories. These models undergo rigorous training and testing processes using binary class datasets for comprehensive evaluation. Traditionally, toxic comments are divided into various classes such as hate,


**Figure 1.** Uploading the toxic comments dataset.

 

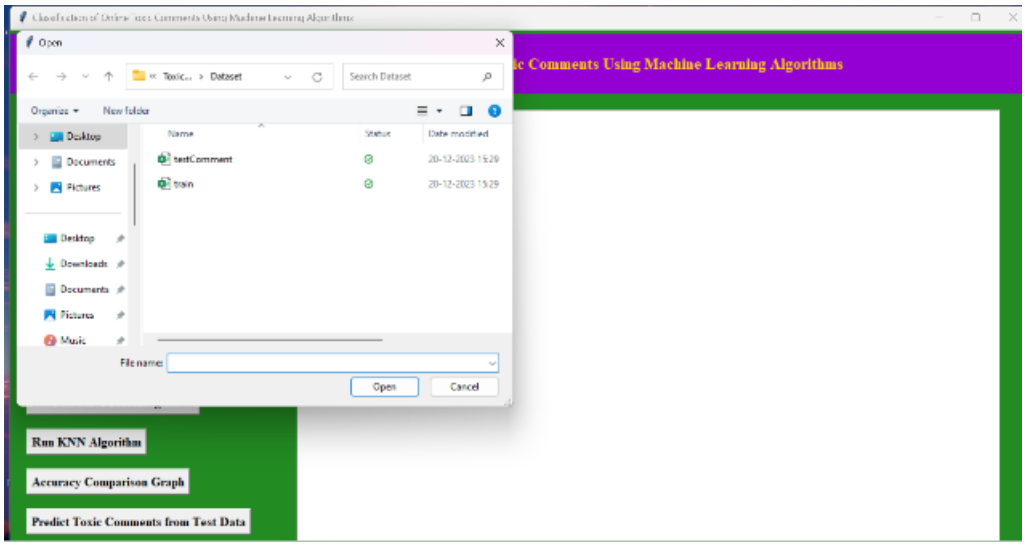


**Figure 2.** Selecting and uploading the 'train.csv' file.

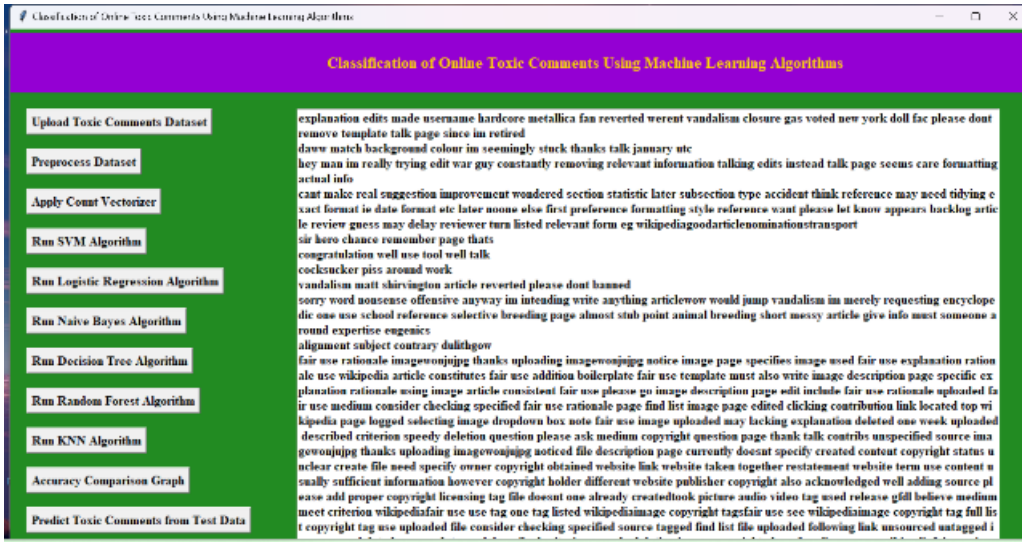


**Figure 3.** Removing stop words and special symbols.

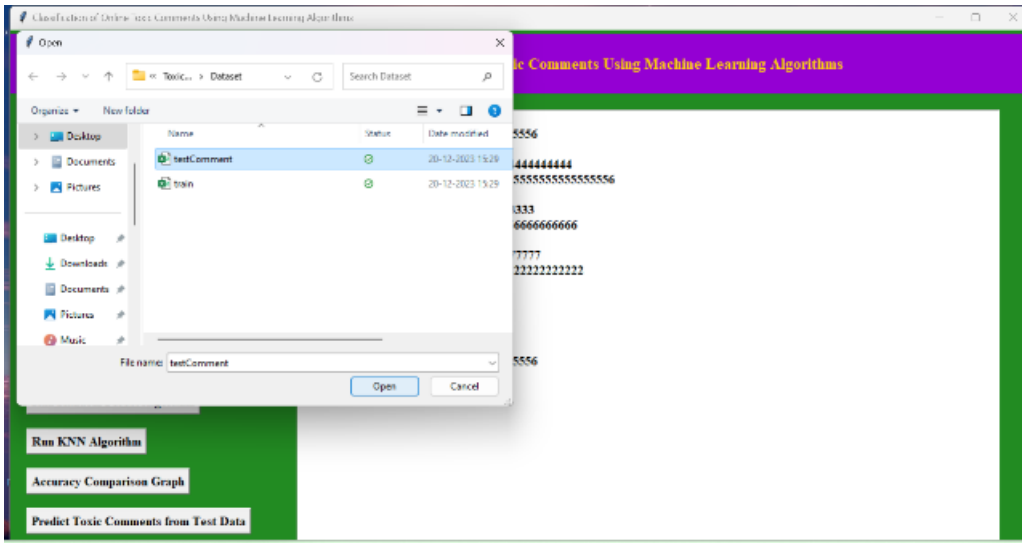


**Figure 4.** Selecting and uploading the 'testComment.csv' file.

 

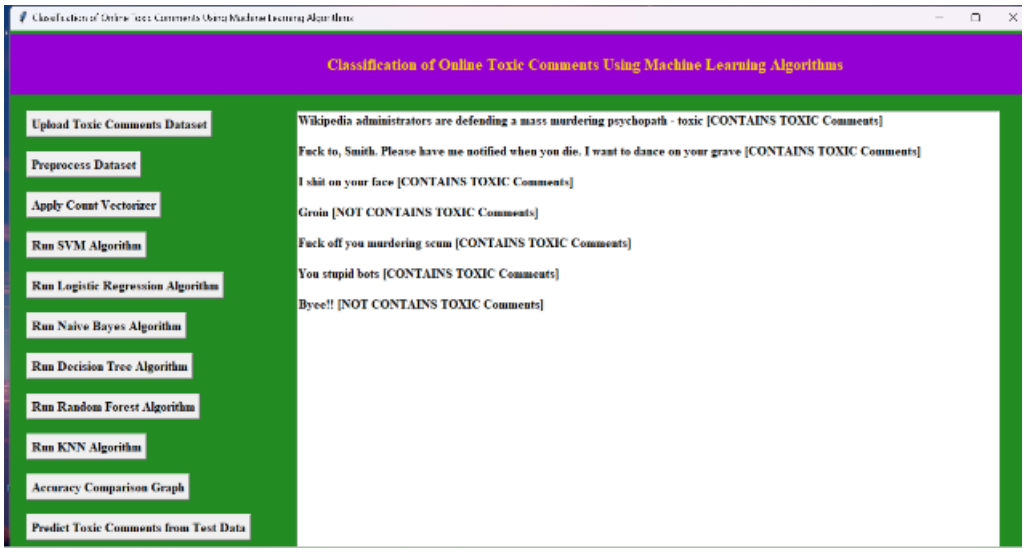


**Figure 5.** Displaying the predicted result.

**Table 1.** Example of variables in the dataset.

| Comment Text | A | B | C | D |
|---|---|---|---|---|
| Toxic | 1 | 0 | 0 | 0 |
| Severe_toxic | 1 | 0 | 1 | 0 |
| Obscene | 0 | 0 | 0 | 0 |
| Threat | 0 | 1 | 0 | 0 |
| Insult | 0 | 0 | 0 | 1 |
| Identity_hate. | 1 | 0 | 0 | 0 |

**Table 2.** Number of comments that fall under toxic and non-toxic comments.

| Category | Number of comments | Test data |
|---|---|---|
| Non-toxic | 143346 | 70000 |
| Toxic | 15294 | 15294 |
| Total | 158640 | 85294 |

toxicity, threat, severe toxicity, obscenity, insult, and identity-based hate. However, the approach diverges from convention by employing a binary classification scheme, categorizing comments solely into toxic and non-toxic groups. The original dataset, obtained from Kaggle, includes multiple labels such as toxic, severe_toxic, obscene, threat, insult, and identity_hate. Non-toxic comments are grouped within a single class, while the remaining comments exclusively bear toxic labels.

For example, Table 1 indicates that ‘A’ and ‘C’ are classified as toxic, while ‘B’ and ‘D’ are categorized as non-toxic. This classification is based on criteria related to toxic, including severe toxic, obscene, threat, insult, and identity hate.

A toxic and non-toxic comment is identified by this technique, and Table 2 presents the total number of comments, the percentage of toxic comments, and the number of non-toxic comments. In addition to the number of comments classified as toxic, it also shows the total number of comments in the dataset as a whole.

This technique discerns between toxic and non-toxic comments, with Table 2 illustrating the aggregate count of comments, the proportion of toxic comments, and the volume of non-toxic comments. Furthermore, it delineates not only the tally of comments classified as toxic but also the comprehensive dataset size.

 

## CONCLUSION

In this study, a supervised learning approach was employed for toxic comment detection on social media platforms. Utilizing a dataset obtained online and meticulously processed to remove noise, various supervised algorithms were applied to classify toxic comments from non-toxic ones. Through rigorous assessment and accuracy measurements, the most proficient models for online toxic comment classification were identified, with their performance visually represented through accuracy graphs. The results indicate that the utilization of machine learning algorithms such as SVM, logistic regression, naive Bayes, random forest, decision tree, and KNN exhibit promise in effectively differentiating toxic comments from non-toxic ones. Furthermore, the integration of a multi-faceted model, which assesses various dimensions of toxicity including severity, obscenity, threat, insult, and identity-hate, significantly bolsters the precision of toxic comment detection.

### Future Scope

Moving forward, there are several avenues for future research and development in the domain of toxic comment detection on social media platforms:

- *Enhanced model optimization:* Further refinement and optimization of machine learning models can enhance the accuracy and efficiency of toxic comment detection. Exploring advanced algorithms and techniques, such as deep learning and ensemble methods, has the potential to yield more robust results.
- *Real-time detection:* The integration of real-time detection capabilities into social media platforms can facilitate immediate action against toxic comments, thereby fostering a safer online environment. This may involve the development of lightweight models optimized for quick inference on streaming data.
- *Multimodal analysis:* Incorporating multimodal analysis, which encompasses textual, visual, and contextual information, holds promise for improving the accuracy and granularity of toxic comment detection. By leveraging additional features such as user metadata and community context, the model can gain a better understanding of the nuanced nature of toxic behavior.

## REFERENCES

1. Badjatiya P, Gupta S, Gupta M, Varma V. Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3–7, 2017. pp. 759–760.
2. Prusa J, Khoshgoftaar TM, Dittman DJ, Napolitano A. Using random undersampling to alleviate class imbalance on tweet sentiment data. In: Proceedings of the IEEE International Conference on Information Reuse and Integration, August 13–15, 2015. pp. 197–202.
3. Waseem Z, Davidson T, Warmsley D, Weber I. A typology of abusive language detection subtasks for understanding abuse. In: Proceedings of the the First Workshop on Abusive Language Online. Vancouver, British Columbia, Canada: Association for Computational Lingistics; 2017. pp. 78–84.
4. Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS. A performance comparison of supervised machine learning models for COVID-19 tweets sentiment analysis. PLoS One. 2021; 16 (2): e0245909.
5. Fatima EB, Omar B, Abdelmajid EM, Rustam F, Mehmood A, Choi GS. Minimizing the overlapping degree to improve class imbalanced learning under sparse feature selection: application to fraud detection. IEEE Access. 2021; 9: 28101–28110.
6. Rustam F, Ashraf I, Mehmood A, Ullah S, Choi GS. Tweets classification on the base of sentiments for US airline companies. Entropy. 2019; 21 (11): 1078.
7. Anandarajan M, Hill C, Nolan T. Practical Text Analytics: Maximizing the Value of Text Data. Advances in Analytics and Data Science, vol. 2. Cham, Switzerland: Springer; 2019.
8. Raja Rao PBV, Prasad M, Kiran Sree P, Venkata Ramana C, Satyanarayana Murty PT. Enhancing the MANET AODV forecast of a broken link with LBP. In: Reddy VS, Prasad VK, Wang J, Rao Dasari NM, editors. Intelligent Systems and Sustainable Computing. ICISSC 2022. Smart Innovation, Systems and Technologies, vol. 363. Singapore: Springer; 2023. pp. 51–69.. doi: 10.1007/978-981-99-4717-1_6.

 

9. Maddula P, Srikanth P, Kiran Sree P, Raja Rao PBV, Satyanarayana Murty PT. COVID-19 prediction with chest X-ray images using CNN. In: 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India, January 27–28, 2023. pp. 568–572. doi: 10.1109/IITCEE57236.2023.10090951.
10. Prasad M, Ajita Lakshmi K, Para Rao PBV, Prasanthi BV, Kiran Sree P, Rajesh Babu V, Das GS. A CNN and TF techniques development for efficient identification of floral recognition. In: 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, February 9–10, 2024. pp. 327–332. doi: 10.1109/IC2PCT60090.2024.10486528.
11. Satyanarayana Murty PT, Prasad M, Raja Rao PBV, Kiran Sree P, Ramesh Babu G, Varma CP. A hybrid intelligent cryptography algorithm for distributed big data storage in cloud computing security. In: Morusupalli R, Dandibhotla TS, Atluri VV, Windridge D, Lingras P, Komati VR, editors. Multi-disciplinary Trends in Artificial Intelligence. MIWAI 2023. Lecture Notes in Computer Science, vol 14078. Cham, Switzerland: Springer; 2023. pp. 637–648. doi: 10.1007/978-3-031-36402-0_59.
12. Kiran Sree P, Chintalapati PV, Usha Devi SSSN, Prasad M, Ramesh Babu G, Raja Rao PBV. Waste management detection using deep learning. In: 2023 3rd International Conference on Computing and Information Technology (ICCIT), Tabuk, Saudi Arabia, September 13–14, 2023. pp. 50–54. doi: 10.1109/ICCIT58132.2023.10273898.

 