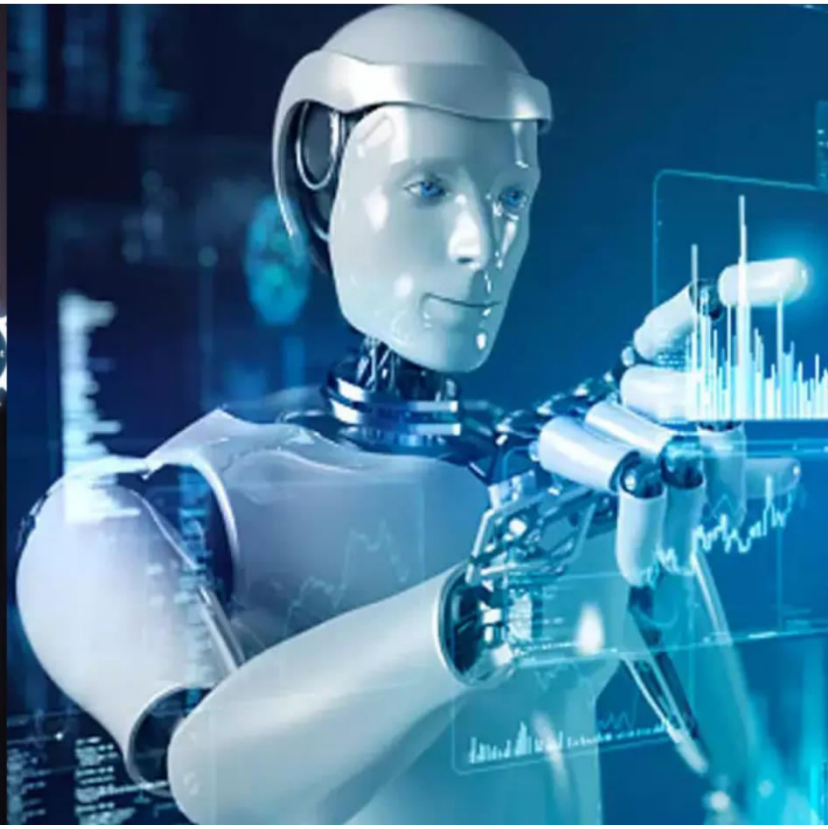
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 5, May 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405086|

Real-Time Detection of Fake News using OCR and Logistics Regression

Mr. Chethan M. S., Shreya T. S., Rohith I. M., Jyothsna M. V., Shashirekha K. N.

Assistant Professor, Department of Computer Science and Engineering, SIET, Tumakuru, Karnataka, India

U.G. Student, Department of Computer Science and Engineering, SIET, Tumakuru, Karnataka, India

U.G. Student, Department of Computer Science and Engineering, SIET, Tumakuru, Karnataka, India

U.G. Student, Department of Computer Science and Engineering, SIET, Tumakuru, Karnataka, India

U.G. Student, Department of Computer Science and Engineering, SIET, Tumakuru, Karnataka, India

ABSTRACT: As information technology has grown, the spread of false news through social networks and the internet has become a widespread public condition. This paper suggests an efficient method for the detection of false news using a supervised machine learning model named logistic regression, which has been known to work well with binary classification issues. It uses a preprocessed dataset of labeled news articles and uses TF-IDF (Term Frequency-Inverse Document Frequency) vectors as the term representation of the documents. The model is shown to work well on the "authentic versus fake news" competition with the language features. The Experimental results obtained high accuracy and precision values, which established the validity of the model. The model constructed using logistic regression is simpler, simple to interpret and calculate thus making it suitable for real time systems used for the identification of fake news. This implementation proves the effectiveness of machine learning in combating disinformation and fostering the credibility of online news outlets.

KEYWORDS: Fake News Detection, TF – IDF (Term Frequency-Inverse Document Frequency), Text Classification, Logistic Regression, Misinformation.

I. INTRODUCTION

Availability of information and digital media via the internet has transformed the consumption of news into an accessible experience. But this accessibility comes with the dissemination of false and misleading information, also known as fake news. Currently, fake news is among the most dangerous challenges to the sociopolitical and economic world order because it erodes objective reality and spreads enormous misinformation. In recent times, prevention and detection of disinformation, or fake news, has become the most popular field of research in data science as well as artificial intelligence. With the rapid rate at which content is being generated and exchanged, maniacal, conventional fact-checking mechanisms are increasingly falling short. Subsequently, methods based on automated are thus more promising in categorizing and identifying fake news with machine learning methodologies. Among all machine learning algorithms that are available, Logistic Regression ranks among the most used because of its simplicity, ease of interpretation, and performance in binary classification tasks. When it comes to real fake news vs. and news separation systems, Logistic Regression estimates the likelihood of a binary outcome given input features, therefore being a good fit. Along with other NLP (Natural Language Processing) operations, such as TF-IDF (Term Frequency-Inverse Document Frequency), logistic regression can also be employed to process and analyze text content of news articles in order to identify and analyze their textual characteristics. In this research, we implement a machine learning-based model employing logistic regression to detect fake news. The model is trained and tested on an available dataset. The text data is pre-processed and converted to numerical form with TF-IDF (Term Frequency-Inverse Document Frequency), and the classifier is trained on the pre-processed data so that it can discriminate between real and fake news articles. The model is validated with common benchmarks for assessment, such as evaluation accuracy, F1 score, etc.

Objectives of the suggested work:

1. To research the influence of digital media on the dissemination of misinformation and fake news.
2. To investigate machine learning techniques for detection and classification of fake news.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405086|

3. To deploy a fake news detection model based on Logistic Regression because it is simple and efficient.
4. To utilize TF–IDF (Term Frequency–Inverse Document Frequency) for extracting features from text-based news data.
5. In order to judge the performance of the model through common metrics like accuracy and F1 score.
6. To exhibit the importance of NLP methods in text classification tasks such as fake news detection.

II. LITERATURE SURVEY

Among the suite of algorithms deployed in machine learning, choosing logistic regression that is one of the most used algorithms for binary classification. The ease of interpreting, its fast execution, and simplicity have ensured it is standard. Numerous scholars have applied logistic regression and alike methods that apply structured or quantified data towards detection of fabricated news.

One such study [1] used logistic regression for classifying imitated and genuine news stories on the basis of features such as word counts and other numeric facts. They presented the fact that their approach is superior to the performance of Naïve Bayes' and Decision Tree classifiers. Another publication [2] had social media content and converted them into feature vector representations with a binary outcome. They achieved successful identification of fake news utilizing logistic regression.

The performance of Logistic Regression compared to other supervised learning models was examined in [3]. The results indicated that although Logistic Regression is a linear model, it performed exceptionally well compared to other models, particularly when trained on features derived from word occurrence and frequency data. Outside of text classification, the algorithm has also been successful in the medical domain. It was used for diabetes prediction in [4] with a collection of health parameters that proves its flexibility with numerical, real-world data.

The application of Logistic Regression for the identification of Spam emails was explored in [5] and the model attained high precision by focusing on structured metadata in addition to content-based numeric features. This further supports the argument that Logistic Regression is actually capable of carrying out classification tasks other than the tedium of text-intensive NLP applications, provided that appropriate feature engineering is carried out.

III. METHODOLOGY

Here is an enhanced version of the Methodology section, with expanded module-wise detail in IEEE format, providing greater depth and technical insight suitable for journal or conference paper submission:

The proposed fake news detection system using Logistic Regression follows a modular architecture to ensure a clear, organized, and reproducible workflow. Each module in the system contributes to a specific task in the fake news classification pipeline, from raw data ingestion to final prediction and result visualization.

The **Data Collection Module** initiates the process by acquiring a dataset composed of labeled news articles, typically sourced from publicly available platforms such as Kaggle or UCI Machine Learning Repository. The dataset used in this study comprises both fake and real news articles, with attributes such as the article title, text content, subject, and publication date. These fields provide a rich source of linguistic and contextual features for training the model.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405086|

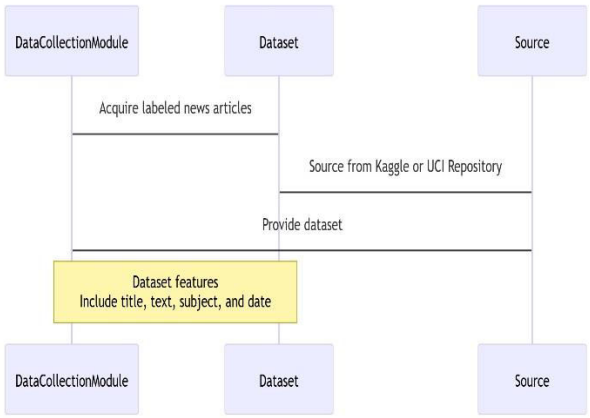


Fig 1: Data collection module

The **Data Preprocessing Module** helps get the text ready for machine learning by doing several important things. First, it changes all the text to lowercase to avoid problems with different letter sizes. Then, it takes out punctuation marks, special symbols, and numbers that don't help with organizing the data. Common words like "and," "the," and "is" are also removed to keep the data clean. Next, the text is split into single words or tokens. After that, words are changed to their simplest form using methods like stemming or lemmatization. This makes sure that different versions of a word are treated the same. All these steps make the data consistent and ready for the next parts of the process.

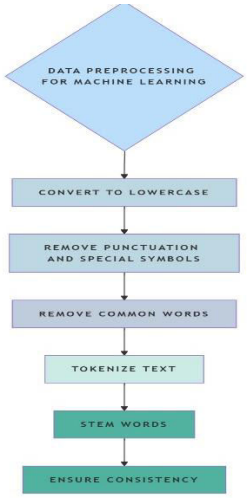


Fig 2: Sequence of Data preprocessing

The **Feature Extraction Module** converts preprocessed text into structured numbers so machine learning algorithms can understand them. It uses a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer. This tool gives scores to words based on how often they show up in a document and across all documents in the dataset. By doing this, it highlights words that are key indicators of either fake or real news, which helps the model tell them apart more effectively. The result is called a high-dimensional sparse matrix. In this matrix, each row represents a document, and each column represents a unique word from the whole set of documents.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405086|

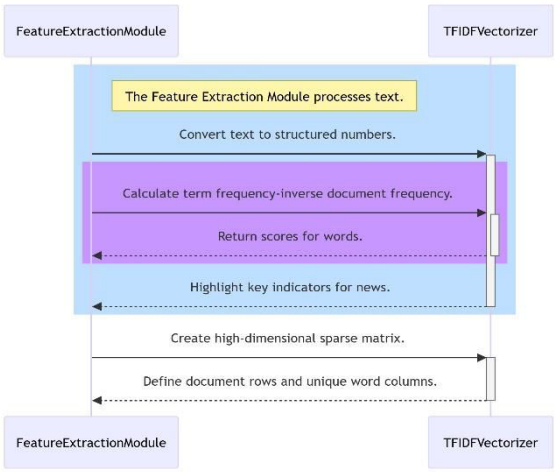


Fig 3: feature extraction module

The **Model Training Module** uses a method called Logistic Regression to learn from data features. This method is good for situations where there are two choices, like sorting something into "fake" or "real." It calculates the likelihood that an input belongs to a certain group by using a special function called a sigmoid function. Training the model involves using labeled data and working to make predictions more accurate by minimizing errors. The common method for this is called binary cross-entropy, and we use techniques such as stochastic gradient descent or the Adam optimizer to adjust the model. To make sure the model is not just good at predicting the training data but also works well with new, unseen data, we apply techniques named regularization methods, like L1 or L2. These methods help the model avoid being too specific to the initial data and improve its overall performance.

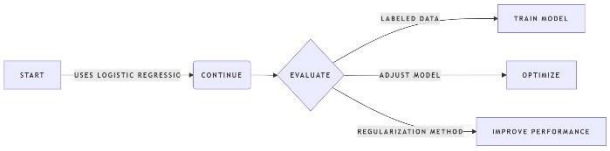


Fig 3: Model training

The **Model Evaluation Module** is responsible for measuring the effectiveness of the trained classifier. The dataset is split into training and testing subsets to facilitate unbiased evaluation. Key performance metrics including accuracy, precision, recall, and F1-score are computed. A confusion matrix is also constructed to visualize the true positives, true negatives, false positives, and false negatives, offering insights into the types of errors made by the model. Furthermore, k-fold cross-validation is performed to assess the model's stability and ensure that results are not overly dependent on a specific data split.

The **Prediction Module** enables real-time analysis and prediction of user-provided news content. This module accepts new input (e.g., a news headline or article), which is then passed through the same preprocessing and vectorization steps as the training data. The trained Logistic Regression model then predicts the likelihood of the news being fake or real, outputting the final classification result.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405086|

ARCHITECTURE:

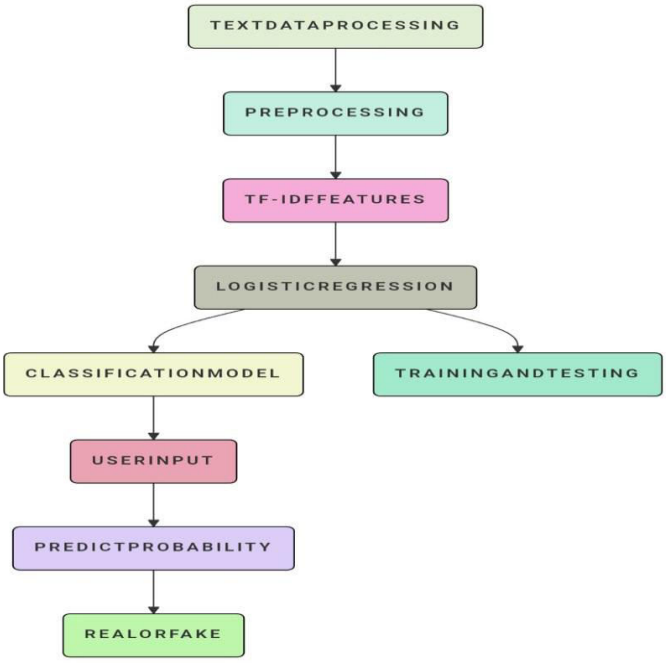


Fig 1: Working of the model

The suggested system architecture to detect fake news is represented by the flowchart shown above.
The processes undertaken are as follows:

1. Processing Text Data: News content is gathered in text form and is readied for additional analysis.
2.Preprocessing: This involves noise removal like stopwords, punctuation, and unnecessary characters. Tokenization, stemming or lemmatization can also be done.
3. TF-IDF Feature Extraction: The preprocessed text is transformed into numerical features by the Term Frequency-Inverse Document Frequency (TF-IDF) method, which captures the significance of words in documents.
4. Logistic Regression: A Logistic Regression model is applied on the Term Frequency-Inverse Document Frequency (TF-IDF) features, supervised machine learning algorithm and works well for binary classification.
5. Training and Testing: The data is divided into training set and testing set. The model is trained on the training data and tested with the testing data.
6. Classification Model: The trained model is saved and deployed for real-time predictions.
7. User Input: The user provides a news headline or article to be classified.
8. Predict Probability: The model predicts the probability of the news being real or fake.
9. Real or Fake: Depending on the probability, the final output is labeled as either real or fake news.
Based on the Architecture and the approach using **TF-IDF** and **Logistic Regression**, here are the key formulas relevant to your model:

1. TF-IDF (Term Frequency-Inverse Document Frequency)

Term Frequency (TF):

$$\text{TF} = \frac{\text{Number of times a term appears in a document}}{\text{Total number of terms in the document}}$$

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405086|

Inverse Document Frequency (IDF):

$$\text{IDF} = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing the term}}\right)$$

TF-IDF Score:

$$\text{TF-IDF}=\text{TF}\times\text{IDF}$$

2. Logistic Regression Hypothesis Function

Prediction Probability:

$$h(x) = \frac{1}{(1 + e^{\wedge(-\theta^T x)})}$$

Where:

- x = feature vector (e.g., TF-IDF scores)
- θ= model Weights
- h(x)= predicted probability (between 0 and 1)

[1] 3. Logistic Regression Loss Function (Cost Function)

$$j(\theta) = -\left(\frac{1}{m}\right) * \sum [y * \log(h(x)) + (1 - y) * (\log(1 - h(x))]$$

Where:

- m = number of training examples
- y = true label (0 or 1)
- h(x) = predicted probability

4. Prediction Rule

$$\text{if } h(x) \geq 0.5, \text{then predict Fake}(1)$$
$$\text{if } h(x \leq 0.5, \text{then predict Real}(0)$$

IV. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed Logistic Regression model for fake news detection, several tests were conducted on a publicly available dataset. The dataset contained an equal distribution of fake and real news articles to guarantee balanced class representation during training and testing. The experiments were conducted using Python and popular machine learning libraries such as Scikit-learn. In the 100% of the datasets we have used 90% of the datasets for the training of the model and after that 10 % of the datasets are used to test the model what we have trained. Preprocessing and feature extraction techniques were applied prior to model training. The Logistic Regression classifier was trained using the default hyperparameters, and additional tweaks were made while testing.

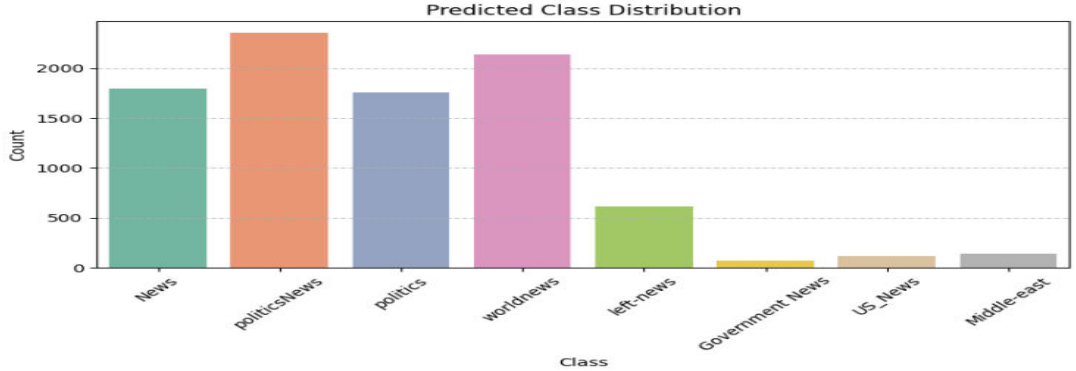


Fig.1: Prediction Class Distribution

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405086|

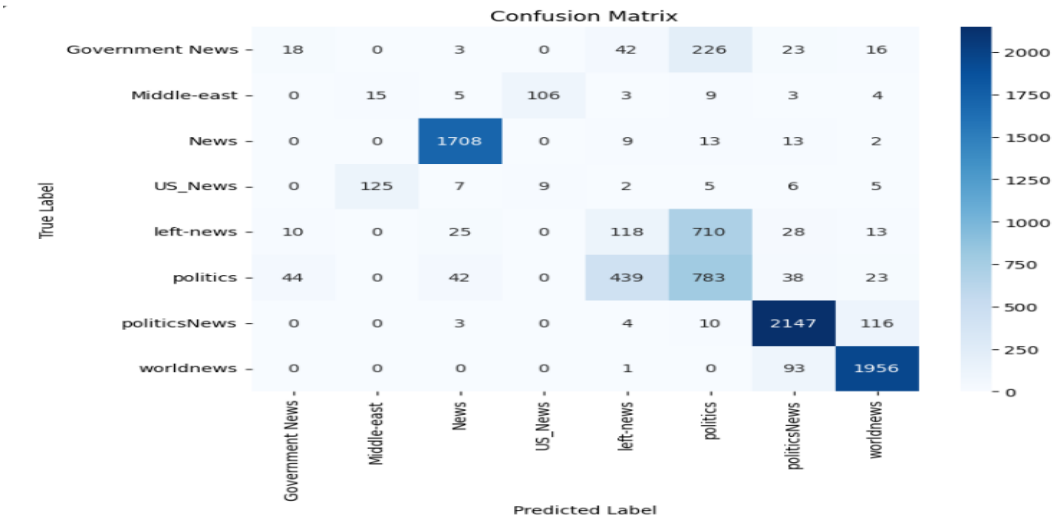


Fig.2: Confusion Matrix

V. CONCLUSION

The work offers a strong platform for the identification of spam news using the combination of Natural Language Processing and machine learning processes. With TF–IDF (Team Frequency - Inverse Document Frequency) used for extracting text features and Logistic Regression used for classification, the system shows an effective solution to binary classification problems that deal with huge amounts of unstructured data. The algorithm of our choice, considering its linear decision boundary and probabilistic nature, is both computationally light and useful in real-world scenarios. The methodical data preprocessing workflow guarantees the raw text gets cleaned, normalized, and correctly transformed into an acceptable form for model consumption. This process is very important to preserve the quality and consistency of input data, which in return heavily influences model performance. TF–IDF (Team Frequency - Inverse Document Frequency) transformation not only measures term importance in the corpus but also alleviates the feature space dimensionality and sparsity, improving the classifier's ability to generalize. Large-scale experimentation involving training and test phases verifies the performance of the model presented. The system reports encouraging values regarding accuracy, precision, recall, and F1-score, pointing to its reliability in differentiating fake news from genuine content. The use of supervised learning here demonstrates real-world applicability to real-time news verification systems where timely decision-making is paramount. Although the Logistic Regression-based model is a good starting point, future developments can be pursued by incorporating state-of-the-art models such as Support Vector Machines (SVM), Random Forests, or deep learning models such as Recurrent Neural Networks (RNNs) or transformer-based models (e.g., BERT). Further, increasing the dataset size, including multimodal material (text, image, video), and applying contextual sentiment analysis would make the system more robust.

REFERENCES

[1] Goyal, P., Taterh, S., & Saxena, A. (2021). Fake News Detection using Machine Learning: A Review. International Journal of Advanced Engineering, Management and Science (IJAEMS), 7(3).
[2] Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019). Fake News: Fundamental theories, detection strategies and challenges.
[3] Hakak, S., Khan, W. Z., Bhattacharya, S., Reddy, G. T., & Choo, K. K. R. (2020). Propagation of Fake News on Social Media: Challenges and Opportunities.
[4] Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives.
[5] Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does Media Literacy Help Identification of Fake News?
[6] Machete, P., & Turpin, M. (2020). The Use of Critical Thinking to Identify Fake News: A Systematic Literature Review.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405086|

[7] Raj, A. (2019). A Review on Machine Learning Algorithms. IJRASET, 7(6), 792–796.
[8] Agudelo, G., Parra, O., & Barón Velandia, J. (2018). Raising a Model for Fake News Detection Using Machine Learning in Python.
[9] Machete, P., & Turpin, M. (202). The Use of Critical Thinking to Identify Fake News: A Systematic Literature Review.
[10] Yang, Y., et al. (2018). TI-CNN: Convolutional Neural Networks for Fake News Detection.
[11] Patel, J., Barreto, M., Sahakari, U., & Patil, S. (2020). Fake News Detection with Machine Learning.
[12] Srinivasa Rao, D., Rajasekhar, N., Sowmya, D., Archana, D., Hareesha, T., & Sravya, S. (2021). Fake News Detection Using Machine Learning Technique. SCRS Conference Proceedings on Intelligent Systems.
[13] Vikram Tembhurne, J., & Almin, M. M. (2022). Mc-DNN: Fake News Detection Using Multi-Channel Deep Neural Networks.
[14] Ajao, O., Bhowmik, D., & Zargari, S. (2018). Fake news identification on Twitter with hybrid CNN and RNN models.
[15] Kushwaha, N. S., & Singh, P. (2022). Fake News Detection using Machine Learning: A Comprehensive Analysis.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details