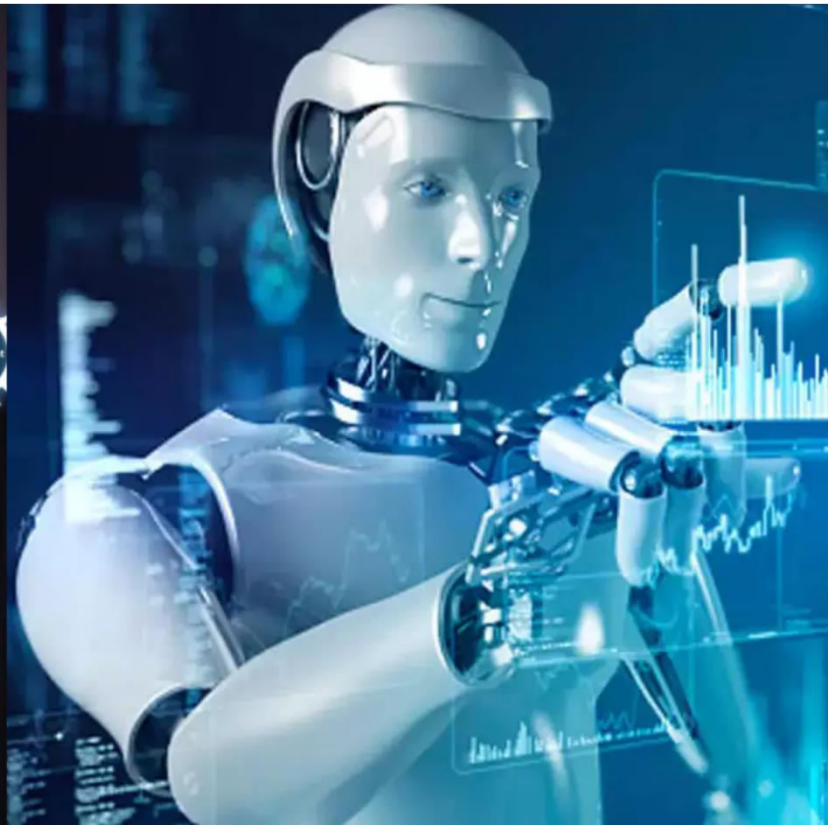
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

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 15, Issue 3, March 2026

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 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503214|

Detection of Fake News Using Machine Learning and Natural Language Processing Algorithms

Santosh Divekar[1], Trupti Jadhav[2], Tanvi Jarande[3], Purva Kadam[4], Siya Kakade[5]

Lecturer, Department of Computer Engineering, AISSMS Polytechnic, Pune, India.[1]

Student, Department of Computer Engineering, AISSMS Polytechnic, Pune, India.[2]

Student, Department of Computer Engineering, AISSMS Polytechnic, Pune, India.[3]

Student, Department of Computer Engineering, AISSMS Polytechnic, Pune, India.[4]

Student, Department of Computer Engineering, AISSMS Polytechnic, Pune, India.[5]

ABSTRACT: The rapid growth of social media and online news platforms has significantly increased the spread of fake and misleading information. Due to the ease of publishing content on the internet, verifying the authenticity of news has become a challenging task. This project presents a machine learning–based fake news detection system using Natural Language Processing (NLP) techniques. The proposed system uses Logistic Regression to classify news articles as real or fake. Textual data is preprocessed using NLP techniques such as tokenization, stop-word removal, lemmatization, and TF-IDF vectorization. The trained model analyzes news content and predicts its authenticity. Experimental results show that Logistic Regression effectively identifies fake news with satisfactory accuracy. The system provides an efficient and scalable solution for automated fake news detection and helps reduce misinformation in digital media.

KEYWORDS: Fake news detection; machine learning; natural language processing; TF-IDF; logistic regression.

I. INTRODUCTION

Fig1. Fake News Detection System

With the widespread use of the internet and social media platforms, the consumption of online news has increased dramatically. While this digital transformation enables faster information sharing, it also facilitates the rapid spread of false and misleading news. Fake news can negatively impact society by creating misinformation, social panic, political instability, and economic disturbances.

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 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503214|

Traditional manual fact-checking methods are time-consuming, subjective, and inefficient when handling large volumes of online data. Therefore, there is a strong need for automated systems that can accurately and efficiently identify fake news. Machine learning and Natural Language Processing (NLP) provide effective techniques for analyzing textual data and extracting meaningful patterns.

In this project, a fake news detection system is developed using NLP techniques and a Logistic Regression machine learning model. Textual features are extracted using TF-IDF vectorization, and the model classifies news articles as real or fake based on their content. The proposed system demonstrates how machine learning can help reduce the spread of misinformation and improve the reliability of digital information sources.

II. LITERATURE REVIEW

Fake news detection has become important due to the rapid spread of misinformation on social media and online news platforms. Machine learning techniques combined with Natural Language Processing (NLP) are widely used to automatically classify news as real or fake.

Ozbay and Alatas [4] proposed a fake news detection system using supervised machine learning algorithms, including Logistic regression, Decision Tree, Naive Bayes, and Support Vector Machine (SVM). Their study demonstrated that Logistic Regression performs effectively for binary classification problems when textual data is preprocessed using NLP techniques such as tokenization and TF-IDF.

Jain et al. [8] developed a fake news detection model using Logistic Regression, Naive Bayes, and SVM. In their work, textual features were extracted using NLP preprocessing methods, and Logistic Regression showed stable performance due to its simplicity and interpretability. The authors concluded that Logistic Regression is suitable for fake news classification because of its low computational complexity.

Henrique et al. [3] focused on machine learning-based fake news detection across multiple platforms. They applied NLP feature extraction methods such as stopword removal and term frequency–inverse document frequency (TF-IDF) with Logistic Regression classifiers. Their results showed that Logistic Regression achieved reliable accuracy across different datasets and languages.

Ahmad et al. [9] employed ensemble machine learning techniques but reported that Logistic Regression performed consistently well when used as a baseline classifier. Their work emphasized that Logistic Regression combined with effective NLP preprocessing yields competitive results for fake news detection.

Jiang et al. [6] analyzed baseline machine learning classifiers for fake news detection and found that Logistic Regression serves as a strong baseline model for comparison with more complex classifiers. Their study highlighted that Logistic Regression provides good performance when trained on balanced and well-preprocessed datasets.

Manzoor et al. [7] reviewed multiple machine learning-based fake news detection techniques and concluded that Logistic Regression is one of the most commonly used and effective algorithms for content-based fake news detection due to its simplicity, speed, and accuracy.

Based on the existing literature, it is evident that **Logistic Regression** combined with NLP techniques such as tokenization, stopword removal, lemmatization, and TF-IDF feature extraction is an effective and widely adopted approach for fake news detection. Therefore, **Logistic Regression** is a suitable choice for developing a reliable fake news detection system.

III. METHODOLOGY

The proposed fake news detection system is designed using a structured methodology that integrates Natural Language Processing (NLP) with traditional machine learning algorithms. The overall workflow of the system consists of data

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 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503214|

collection, data preprocessing, feature extraction, model training, prediction, and performance evaluation. Each stage plays a critical role in ensuring the accuracy, reliability, and robustness of the fake news classification system.

A. Data Collection

The dataset used in this study was obtained from an open-source fake news dataset available on the Kaggle platform. The dataset consists of a large collection of news articles that are labeled as either *real* or *fake*. Each data record includes textual attributes such as the news article title, author name, and the complete content of the article.

This dataset was selected due to its diversity, real-world relevance, and frequent use in fake news detection research. The presence of labeled data enables supervised machine learning algorithms to learn meaningful patterns and relationships from historical examples, thereby improving prediction accuracy.

B. Data Preprocessing

Textual data collected from online sources often contains noise, inconsistencies, and irrelevant information. Therefore, preprocessing is a crucial step to enhance the quality of the dataset and improve model performance. The following preprocessing techniques were applied:

1) Removal of Punctuation and Special Characters:

Regular expressions were used to remove punctuation marks, numbers, and special symbols that do not contribute to the semantic meaning of the text.

2)Tokenization:

The text was broken down into individual tokens or words, allowing machine learning models to analyze each term independently.

3)Stop-Word Removal:

Commonly occurring words such as *is*, *the*, *and*, and *of* were removed using predefined stop-word lists. This step reduces dimensionality and eliminates unnecessary noise from the dataset.

4)Lemmatization:

Words were converted into their base or root forms (for example, *running* is converted to *run*). This process helps normalize different grammatical forms of the same word.

5)Case Normalization:

All textual data was converted into lowercase to ensure uniformity and consistency across the dataset.

These preprocessing steps significantly improved the quality of textual data and enhanced the effectiveness of the feature extraction process.

C. Feature Extraction

After preprocessing, the textual data must be converted into numerical form so that it can be processed by machine learning algorithms. In this study, the **Term Frequency–Inverse Document Frequency (TF-IDF)** technique was used for feature extraction.

TF-IDF assigns weights to words based on their frequency within a document and their rarity across the entire dataset. Words that occur frequently in a particular document but less frequently across other documents receive higher importance. This method highlights informative terms while reducing the influence of commonly occurring but less meaningful words.

The TF-IDF technique results in a high-dimensional feature space, which is well suited for text classification tasks using traditional machine learning algorithms.

D. Machine Learning Model Development

The extracted TF-IDF feature vectors were used to train a **Logistic Regression classifier**. Logistic Regression was selected due to its efficiency, simplicity, and suitability for binary text classification problems such as fake news detection.

The model estimates the probability of a news article being real or fake using a sigmoid function. Logistic Regression performs well with high-dimensional and sparse text data generated by TF-IDF, making it an appropriate choice for this system.

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 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503214|

The dataset was divided into training and testing subsets using an 80:20 ratio to evaluate the performance and generalization capability of the model.

E. System Architecture

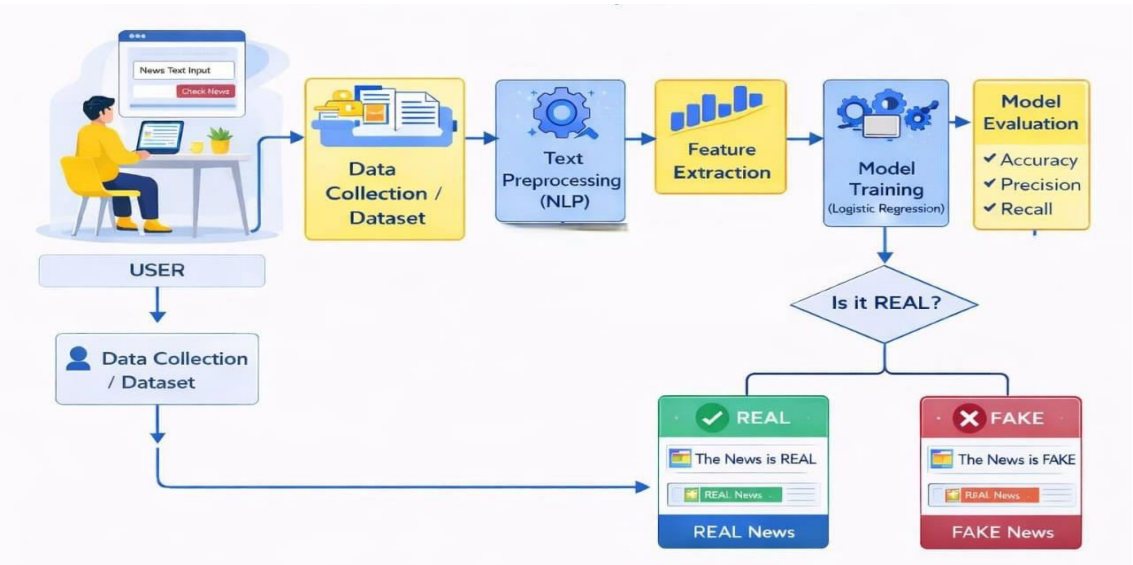


Fig 2. System Architecture

The fake news detection system follows a structured workflow:

1. Input news article
2. Text preprocessing using NLP techniques
3. Feature extraction using TF-IDF
4. Classification using trained ML model
5. Output prediction (Real or Fake)

These evaluation metrics provide a comprehensive assessment of the effectiveness, reliability, and predictive capability of the proposed fake news detection system.

IV RESULT AND DISCUSSION

A. Model Performance

The performance of the fake news detection system was evaluated using accuracy as the primary metric. The Logistic Regression model achieved an accuracy of 98.45% in classifying news articles as real or fake. This indicates that the model is capable of correctly identifying a high proportion of news articles in the dataset.

B. Discussion

Logistic Regression provided stable and interpretable results due to its linear nature and probabilistic output. Its fast training time and low computational complexity make it suitable for fake news detection tasks. The experimental results confirm that Logistic Regression is an efficient and effective model for binary text classification problems when combined with NLP preprocessing techniques such as tokenization, stopword removal, lemmatization, and TF-IDF vectorization.

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

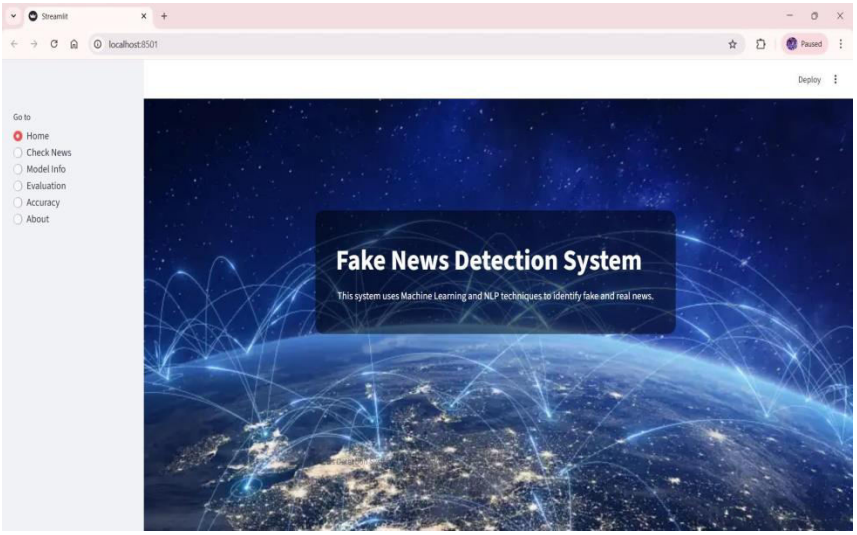
Impact Factor 8.699

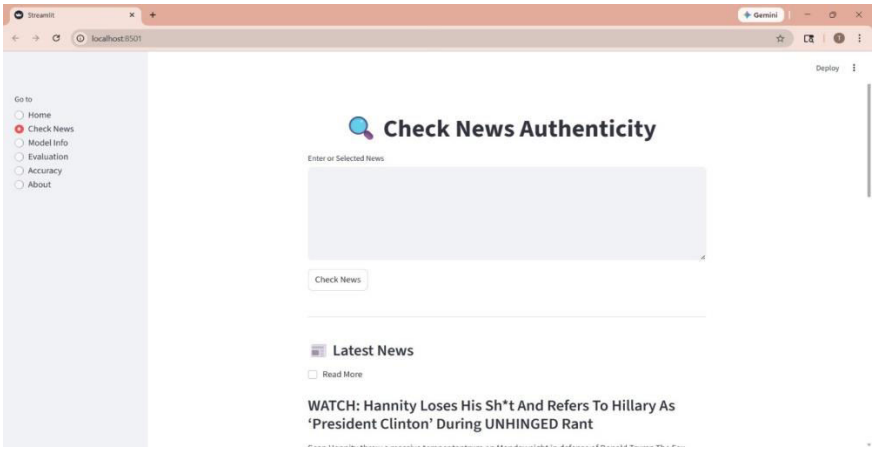
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

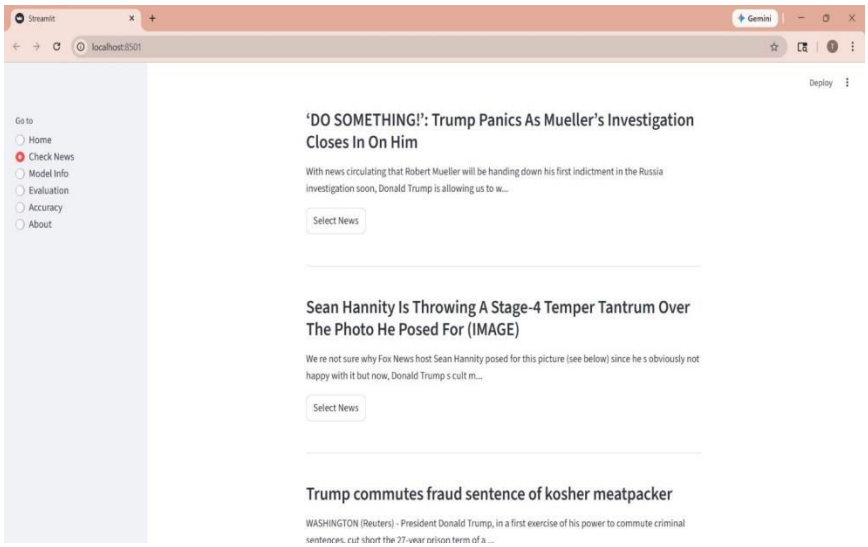
Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503214|

V. RESULT







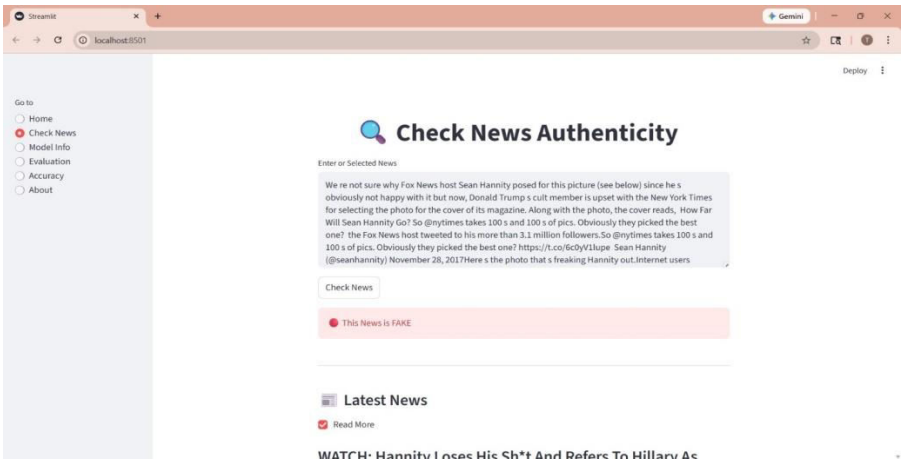
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

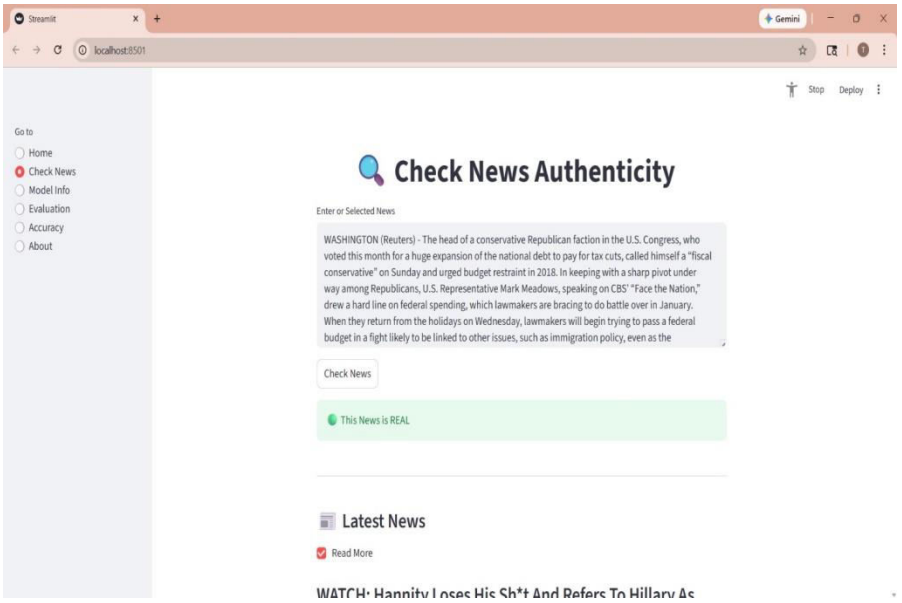
Impact Factor 8.699

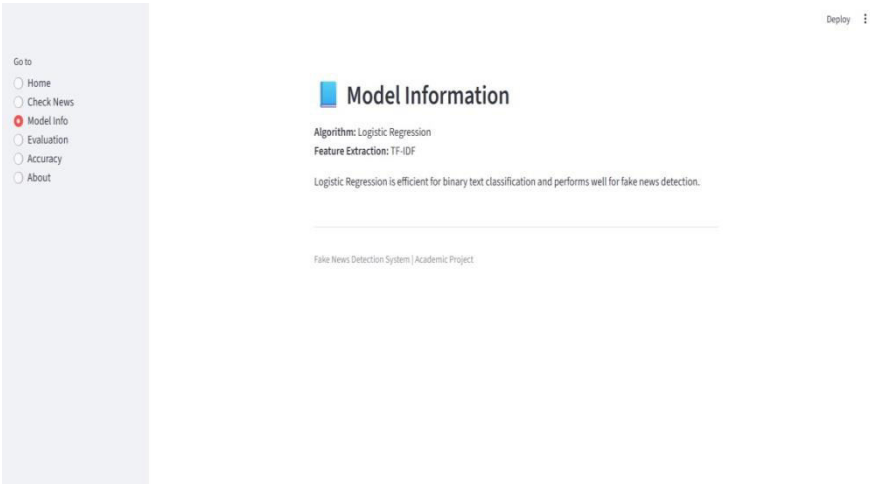
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503214|







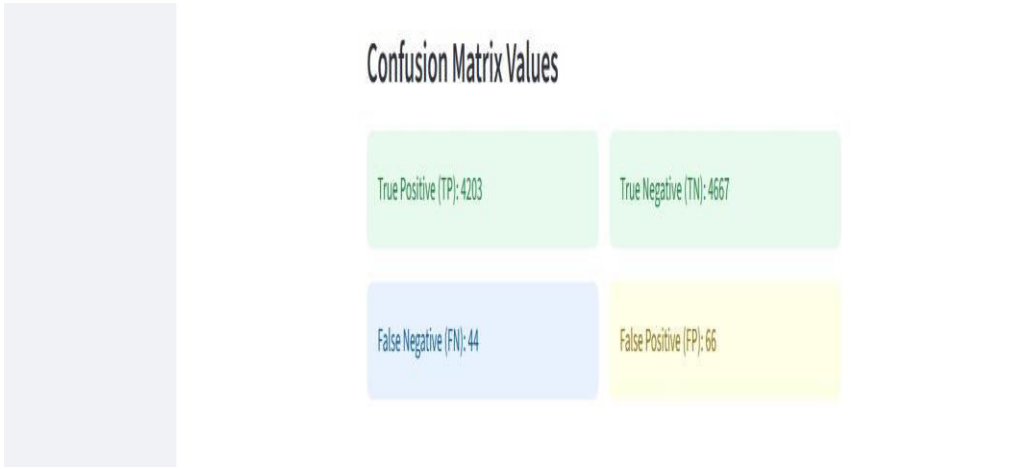
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

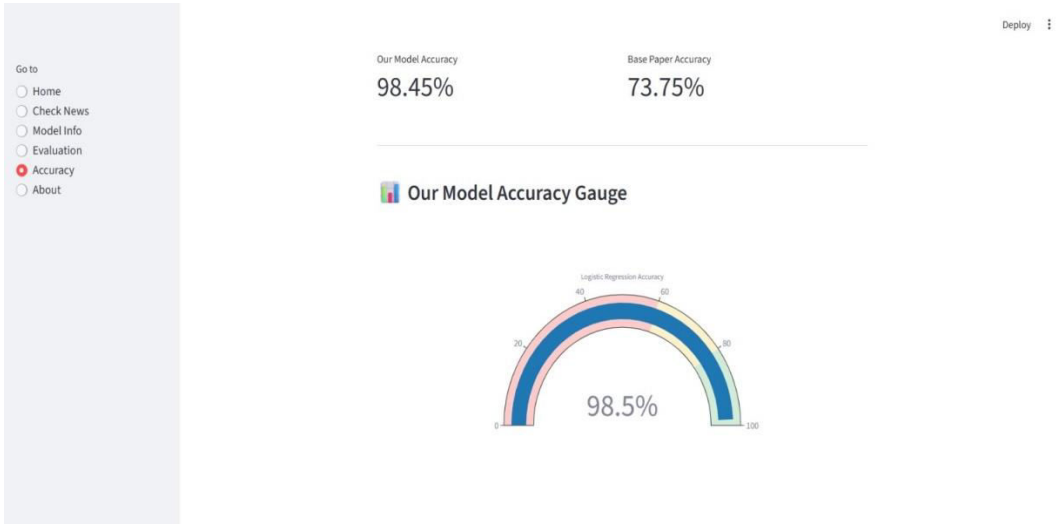
Impact Factor 8.699

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

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503214|





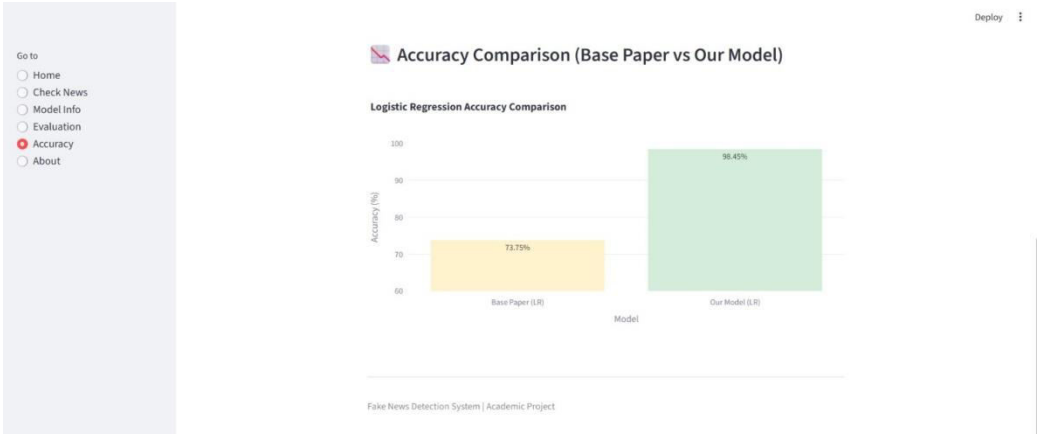
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

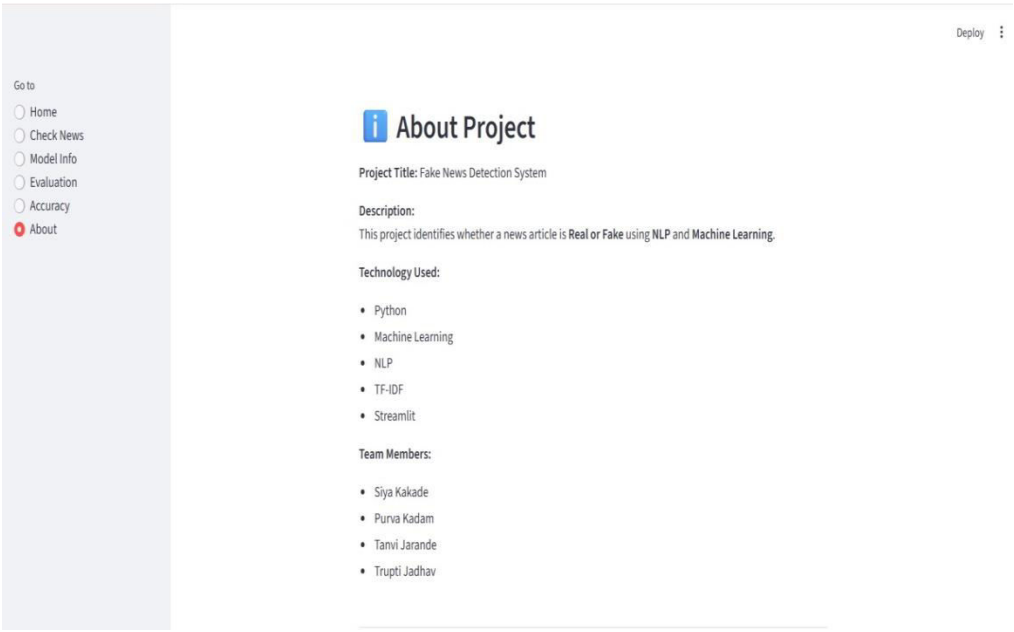
Impact Factor 8.699

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

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503214|





VI. FUTURE SCOPE

- The system can be enhanced using advanced Machine Learning and Deep Learning models to improve detection accuracy and performance.
- The project can be extended to detect fake news from images and videos along with textual content.
- Multi-language support can be added to enable the system to analyze news in different regional and global languages.
- The system can be integrated with websites or browser extensions to provide instant fake news alerts to users.
- A real-time data processing feature can be implemented to analyze live news feeds and social media content.
- User feedback and reporting mechanisms can be added to continuously improve the model's accuracy and reliability.

VII. CONCLUSION

Fake news is a major challenge in today's digital world, affecting people's trust, awareness, and decision-making. With the rise of social media and online platforms, misinformation spreads quickly, making it difficult to distinguish between real and 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 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503214|

This project demonstrates how Machine Learning and Natural Language Processing (NLP) can be used to classify news as real or fake. By analyzing text data such as headlines and content, the system identifies patterns that help in accurate classification.

The developed system is simple, efficient, and helps reduce the spread of misinformation by allowing users to verify news before trusting or sharing it.

REFERENCES

[1] J. Strömbäck, Y. Tsfati, H. Boomgaarden, et al., "News media trust and its impact on media use: Toward a framework for future research," Annals of the International Communication Association,
vol. 44, pp. 139-156, 2020.
[2] E. Mitchelstein and P. J. Boczkowski, "Online news consumption research: An assessment of past work and an agenda for the future," New Media & Society, vol. 12, pp. 1085-1102, 2010.
[3] P. Henrique, A. Faustini and T. F. Covões, "Fake news detection in multiple platforms and languages," Expert Systems with Applications, vol. 158, pp. 1-9, 2020.
[4] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms,"
Physica A: Statistical Mechanics and its Applications, vol. 540, pp. 1-19, 2020.
[5] M. Umer, "Fake news stance detection using deep learning architecture (CNN-LSTM)," IEEE Access, vol. 8, pp. 156695-
156706, 2020.
[6] T. Jiang, J. P. Li, A. U. Haq, et al., "A novel stacking approach for accurate detection of fake news," IEEE Access, vol. 9, pp. 22626-22639, 2021.
[7] S. I. Manzoor, J. Singla, and Nikita,"Fake news detection using machine learning approaches: A systematic review," in Proc.
International Conference on Trends in Electronics and Informatics, 2019, pp. 230-234.
[8] A. Jain, A. Shakya, H. Khatter, et al., "A smart system for fake news detection using machine learning," in Proc. International
Conference on Issues and Challenges in Intelligent Computing Techniques, 2019, pp. 1-4.
[9] I. Ahmad, M. Yousaf, S. Yousaf, et al., "Fake news detection using machine learning ensemble methods," Complexity, pp. 1-11, 2020.
[10] UTK machine learning club. (July 2017). Fake news, version 1. [Online]. Available: https://www.kaggle.com/c/fake-news/data
[11] H. Ali, M. S. Khan, A. AlGhadhban, et al., "All your fake detector are belong to us: Evaluating adversarial robustness of fake-news detectors under black-box settings," IEEE Access, vol. 9, pp. 81678-81692, 2021.
[12] I. K, Sastrawan, I. P. A. Bayupati, and D. M. S. Arsa, "Detection of fake news using deep learning CNN-RNN based methods," ICT Express, pp. 1-13, 2021.
[13] Y. A. Solangi, Z. A. Solangi, S. Aarain, et al., "Review on Natural Language Processing (NLP) and its toolkits for opinion mining and
sentiment analysis," in Proc. International Conference on Engineering Technologies and Applied Sciences, 2018, pp. 1-4.
[14] G. Kim and S. H. Lee, "Comparison of Korean preprocessing performance according to Tokenizer in NMT transformer model," Journal of Advances in Information Technology, vol. 11, pp. 228-232, 2020.
[15] T. Daghistani and R. Alshammari, "Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes," Journal of Advances in Information Technology, vol. 11, pp. 78-83, 2020.
[16] W. He, Y. He, B. Li, et al., "A naive-Bayes-based fault diagnosis approach for analog circuit by using image-oriented feature extraction and selection technique," IEEE Access, vol. 8, pp. 5065-5079, 2020.

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