© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717884

Explainable AI-Based Fake News Detection System for Indian Online News Using Machine Learning and Deep Learning

ROSHNI PRIYADARSHINI BEHERA[1], RAKSHITHA B.S[2]
[1]*PG Research Scholar, Jain (Deemed-to-be University) Bangalore, India*
[2]*Assistant Professor, Department of Computer Science and Engineering Jain (Deemed to-be University), Bangalore, India*

Abstract- ***The rapid growth of digital media and social net-working platforms has significantly increased the spread of fake news and online misinformation. Social media applications such as WhatsApp, Facebook, Instagram, and X allow information to spread rapidly among users without proper verification. In India, fake news has become a major challenge affecting politics, healthcare, education, finance, and social harmony. This paper presents an Explainable Artificial Intelligence (XAI)-based fake news detection system specifically designed for Indian online news environments. The proposed framework integrates Natural Language Processing (NLP), TF-IDF feature extraction, Multinomial Naive Bayes classification, Long Short-Term Memory (LSTM) deep learning, and Local Interpretable Model-Agnostic Explanations (LIME). Multiple datasets including Fake.csv, True.csv, IFND, Bharat-FakeNewsKosh, and Indian news headline datasets were used to improve contextual relevance for Indian misinformation detec-tion. The system was implemented using Python, Flask, SQLite, TensorFlow, Keras, Scikit-learn, and NLTK. Experimental results demonstrate strong classification perfor-mance and improved explainability through LIME-based inter-pretation. The proposed framework contributes toward trust-worthy, interpretable, and practical AI-based misinformation detection systems for public awareness.***

Index Terms—Fake News Detection, Explainable AI, Machine Learning, Deep Learning, LSTM, LIME, TF-IDF, Naive Bayes, Flask, Indian Misinformation

I. INTRODUCTION

The rapid advancement of internet technologies and digital communication platforms has significantly transformed the way people consume and share information. Social media ap-plications such as WhatsApp, Facebook, Instagram, YouTube, and X (formerly Twitter) have become major sources of news and public communication. In India, the increasing availabil-ity of affordable smartphones and internet connectivity has accelerated the growth of digital information sharing among millions of users.

Although digital media platforms improve accessibility to information, they have also increased the spread of fake news and online misinformation. Fake news refers to false, misleading, manipulated, or intentionally fabricated informa-tion created to deceive readers. Such misinformation spreads rapidly through social media forwarding, online groups, blogs, and video-sharing platforms without proper verification mech-anisms.

In recent years, fake news has become a serious social and political issue in India. Misinformation related to elections, religion, healthcare, finance, and government schemes has created public confusion and social unrest. During health-care emergencies, false medical information can create panic among citizens. Similarly, fake political narratives and com-munal misinformation can influence public opinion and dam-age social harmony.

One of the major reasons for the rapid spread of fake news is the forwarding behavior of users on social media platforms. Messages shared through WhatsApp groups and social media channels are often forwarded repeatedly without verifying authenticity. Fake government schemes, Aadhaar-related scams, and financial fraud messages are common examples of misinformation frequently observed in Indian digital environments.

Traditional manual fact-checking approaches are insufficient for handling the enormous volume of online content generated daily. Human verification processes require significant time and resources, making them ineffective for real-time misinfor-mation detection. Therefore, automated fake news detection systems using Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) have become important re-search areas.

Machine learning algorithms such as Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machines have been widely used for fake news classification. These ap-proaches use Natural Language Processing (NLP) techniques such as TF-IDF and Bag-of-Words for feature extraction. Deep learning approaches such as CNN, RNN, and LSTM further improve semantic understanding and sequential learning capa-bilities.

Although these techniques achieve strong classification per-formance, many existing systems behave as black-box models. Users often receive predictions without understanding why a particular news article was classified as fake or real. This lack of transparency reduces trust and interpretability in AI-based systems.

Explainable Artificial Intelligence (XAI) addresses this limitation by improving transparency and interpretability in machine learning and deep learning systems. Explainabil-ity techniques such as Local Interpretable Model-Agnostic Explanations (LIME) identify important words and features contributing to predictions, allowing users to better understand model behavior.

This paper presents an Explainable AI-based fake news detection system specifically designed for Indian online news environments. The proposed framework integrates Natural Language Processing, TF-IDF feature extraction, Multinomial Naive Bayes classification, Long Short-Term Memory (LSTM) deep learning, LIME explainability, Flask deployment, and SQLite database integration.

Multiple datasets including Fake.csv, True.csv, IFND, BharatFakeNewsKosh, and Indian news headline datasets were integrated to improve contextual understanding of Indian mis-information patterns. The proposed system aims to provide an accurate, transparent, practical, and explainable fake news detection framework for public awareness and misinformation analysis.

II. RELATED WORK

Several researchers have explored fake news detection using machine learning, deep learning, explainable AI, and hy-brid computational approaches. Existing fake news detection systems can be broadly categorized into traditional machine learning approaches, deep learning approaches, explainable AI systems, and advanced hybrid frameworks.

A. Traditional Machine Learning Approaches

Traditional machine learning approaches have been widely used for fake news detection because of their computational ef-ficiency and ease of implementation. These systems generally use textual feature extraction techniques such as Bag-of-Words and TF-IDF combined with machine learning classifiers.

Naive Bayes is one of the most commonly used algorithms for fake news classification because of its probabilistic learning mechanism and efficiency in handling textual data. Logistic Regression and Support Vector Machines (SVM) are also frequently used because of their ability to generate strong classification boundaries.

TF-IDF feature extraction is commonly integrated with traditional machine learning approaches to convert textual data into numerical vectors. These models perform well for short-text classification and sparse datasets.

Although traditional machine learning methods are com-putationally efficient and easier to deploy, they often fail to capture semantic relationships and long-range contextual dependencies in textual information.

B. Deep Learning Approaches

Deep learning approaches have significantly improved fake news detection performance by enabling contextual under-standing and sequential learning capabilities.

Convolutional Neural Networks (CNN) are commonly used for extracting local textual patterns and semantic features from news articles. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks are highly effective for sequential text processing because they can retain contextual information from previous words and sentences.

LSTM networks address the vanishing gradient problem present in traditional RNN models and improve long-range dependency learning. Bidirectional LSTM architectures fur-ther improve contextual understanding by processing textual sequences in both forward and backward directions.

Recent research has also focused on transformer-based architectures such as BERT and multilingual BERT. Trans-former models use attention mechanisms to improve semantic representation and contextual learning. Although these models achieve high accuracy, they require significant computational resources and large-scale datasets for training.

C. Explainable AI Approaches

Explainable Artificial Intelligence has become an important research area because many machine learning and deep learn-ing models behave as black-box systems.

LIME and SHAP are widely used explainability techniques that improve transparency and interpretability. LIME generates local explanations for individual predictions by identifying important features contributing to model decisions. SHAP uses cooperative game theory concepts to calculate feature importance values.

Explainability is particularly important in fake news detec-tion systems because users should understand why a particular article is classified as fake or real. Transparent predictions improve user trust and increase system reliability.

Despite improvements in explainable AI, many existing fake news detection systems still provide limited interpretability and lack practical deployment support.

D. Hybrid and Advanced Approaches

Several researchers have proposed advanced hybrid systems integrating multiple computational techniques for misinforma-tion detection.

Graph Neural Networks (GNN) analyze user interaction networks and misinformation propagation patterns across so-cial media platforms. These approaches model relationships between users, posts, and information-sharing behavior.

Blockchain-based fake news detection frameworks focus on content verification, traceability, and decentralized trust man-agement. Blockchain systems aim to reduce misinformation by validating content authenticity and preventing tampering.

Multimodal fake news detection systems integrate textual, visual, and contextual information to improve classification performance. These systems analyze both textual content and associated images or videos.

Although hybrid approaches improve performance, they often introduce higher computational complexity and deploy-ment challenges. Many advanced systems also lack explain-ability and practical real-time deployment support.

The limitations identified in existing research motivated the development of the proposed explainable AI-based fake news detection framework specifically adapted for Indian online misinformation environments.

III. RESEARCH GAPS

Although significant progress has been made in fake news detection using Artificial Intelligence, Machine Learning, and Deep Learning techniques, several important limitations still exist in current systems.

One major limitation is the lack of explainability in many fake news detection models. Most advanced deep learning systems behave as black-box models, where users receive only the final prediction without understanding the reasoning behind the decision. This lack of transparency reduces trust and makes the systems difficult to interpret in real-world applications.

Another important challenge is the limited availability of Indian-context fake news datasets. Many existing datasets primarily focus on Western news articles and international political narratives. However, misinformation patterns in India are often different and commonly include WhatsApp for-wards, fake government schemes, Aadhaar scams, communal misinformation, manipulated political narratives, and financial fraud messages. Existing models trained only on international datasets may fail to effectively capture these Indian-specific misinformation patterns.

Dataset imbalance is another major issue in fake news detection research. Real-news datasets are usually much larger than fake-news datasets, causing machine learning models to become biased toward predicting real news more frequently. This reduces fake-news sensitivity and increases false-negative predictions.

Computational complexity also remains a challenge in ad-vanced deep learning approaches. Transformer-based archi-tectures such as BERT and GPT-based models require high computational resources, GPU acceleration, and large-scale datasets for training. Such requirements make lightweight deployment difficult for practical public-awareness systems.

Another limitation observed in many existing systems is the absence of real-time deployment support. Several research works focus only on model training and evaluation with-out implementing practical web applications or deployment frameworks for public use. Real-time prediction systems are important for improving accessibility and usability.

Existing fake news detection systems also provide limited support for explainable predictions. Users are often unable to identify which words or phrases influenced the prediction. This becomes especially important in sensitive domains such as healthcare, politics, and government-related misinformation where transparency is essential.

Additionally, many current systems rely primarily on textual analysis and do not effectively combine contextual informa-tion, user behavior, and explainability into a unified frame-work. The integration of explainable AI, real-time deployment, database support, and Indian-context datasets remains limited in current research.

To address these limitations, the proposed framework inte-grates:

- Indian-context fake news datasets
- Dataset balancing techniques
- TF-IDF and Naive Bayes classification
- LSTM deep learning
- Explainable AI using LIME
- Flask-based real-time deployment
- SQLite database integration

The proposed system aims to provide a transparent, in-terpretable, practical, and context-aware fake news detection framework specifically designed for Indian online news envi-ronments.

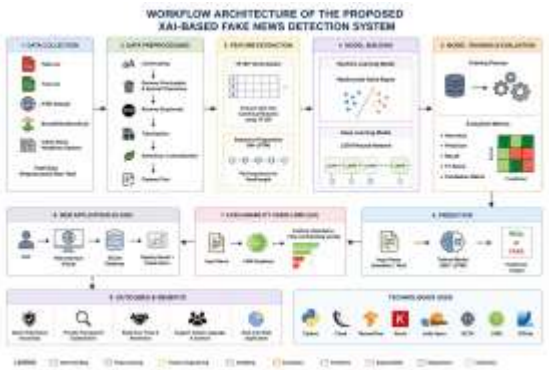

Fig. 1. workflow diagram

IV. METHODOLOGY

The proposed fake news detection framework consists of multiple stages including dataset collection, preprocessing, feature extraction, machine learning classification, deep learn-ing analysis, explainable AI integration, and deployment. The overall objective of the system is to develop an accurate, transparent, and explainable fake news detection framework specifically adapted for Indian online news environments.

The workflow of the proposed system is shown in Fig. 1. The framework begins with dataset collection and preprocess-ing, followed by TF-IDF feature extraction, machine learning and deep learning

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717884

classification, explainability generation us-ing LIME, and Flask-based deployment for real-time predic-tion.

A. Data Collection and Dataset Integration

The quality and diversity of datasets play an important role in fake news detection performance. To improve contextual relevance and generalization capability, multiple datasets were collected and integrated into the proposed framework.

The datasets used in this research include:

- Fake.csv
- True.csv

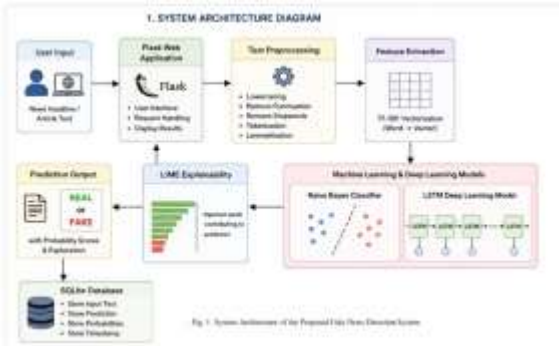

Fig. 2. System Architecture of the Proposed Fake News Detection Framework

- IFND Dataset
- BharatFakeNewsKosh
- Indian News Headlines Dataset

The Fake.csv and True.csv datasets contain large collec-tions of international fake and real news articles. However, fake news patterns in India differ significantly from Western misinformation trends. Therefore, Indian-context datasets such as IFND and BharatFakeNewsKosh were integrated into the system.

The Indian fake-news datasets include examples related to:

- WhatsApp misinformation
- fake government schemes
- Aadhaar scams
- political misinformation
- communal misinformation
- financial fraud messages

Integrating multiple datasets improves contextual under-standing and increases the system's ability to detect Indian online misinformation patterns more effectively.

B. Dataset Balancing

Dataset imbalance is one of the major challenges in fake news detection systems. In many cases, real-news datasets are significantly larger than fake-news datasets. This causes machine learning models to become biased toward predicting real news more frequently.

To reduce prediction bias and improve fake-news sensitivity, dataset balancing techniques were applied. Balanced datasets improve model learning and reduce false-negative predictions. Random sampling and preprocessing techniques were used to maintain approximately equal distributions of fake and real news samples.

The balancing process significantly improved the system's capability to identify fake-news patterns in Indian misinfor-mation scenarios.

C. Text Preprocessing

Text preprocessing is an important step in Natural Language Processing because raw textual data often contains noise, punctuation, special symbols, repeated words, and stopwords that reduce model performance.

The preprocessing pipeline implemented in the proposed system includes:

1) Lowercase conversion
2) Punctuation removal
3) Stopword removal
4) Tokenization
5) Text cleaning
6) Word normalization

Lowercase conversion ensures that words with different capitalization styles are treated uniformly. Punctuation removal eliminates unnecessary symbols that do not contribute signif-icantly to classification.

Stopword removal was performed using the Natural Lan-guage Toolkit (NLTK). Common words such as “the”, “is”, “are”, and “was” were removed because they appear frequently in both fake and real news articles and provide limited semantic value.

Tokenization divides text into smaller units such as words or tokens, making it easier for machine learning models to process textual information. Text cleaning and normalization improve dataset consistency and reduce computational com-plexity.
The preprocessing stage significantly improves feature qual-ity and classification performance.

D. TF-IDF Feature Extraction

Machine learning algorithms cannot directly process raw textual data. Therefore, textual information must be converted into numerical representations before classification.

The proposed system uses Term Frequency–Inverse Doc-ument Frequency (TF-IDF) vectorization to transform news articles into numerical feature vectors.

TF-IDF measures the importance of a word in a document relative to the entire dataset. Frequently occurring words within a document receive higher scores, while common words appearing across many documents receive lower scores.

The TF-IDF formula is represented as:

$$TFIDF(t, d) = TF(t, d) \times log \frac{N}{DF(t)} \quad (1)$$

where:

- TF = Term Frequency
- DF = Document Frequency
- N = Total number of documents

TF-IDF generates sparse matrix representations that effi-ciently capture important textual patterns while reducing the influence of irrelevant words.

The vectorized features generated through TF-IDF were used as inputs for the Naive Bayes machine learning classifier.

E. Naive Bayes Machine Learning Classification

The proposed framework uses a Multinomial Naive Bayes classifier for fake news classification. Naive Bayes is widely used in text classification because of its computational effi-ciency, probabilistic learning capability, and effectiveness for sparse textual data.

The classifier calculates conditional probabilities based on word occurrences within fake and real news categories. The probability-based learning mechanism makes Naive Bayes suitable for textual analysis tasks such as spam filtering, sentiment analysis, and fake news detection.

The major advantages of Naive Bayes include:

- Fast training and prediction
- Low computational complexity
- Effective performance on textual data
- Efficient handling of sparse matrices
- Probability-based classification

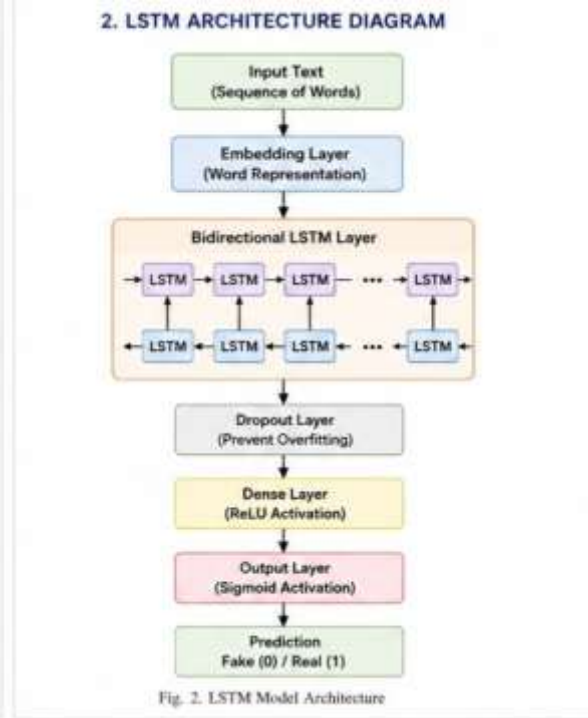

Fig. 3. LSTM Deep Learning Architecture

The trained model generates prediction probabilities indi-cating whether a given news article belongs to the fake or real category.

F. LSTM Deep Learning Classification

Although traditional machine learning models provide effi-cient performance, they often fail to capture long-range con-textual dependencies in textual data. To address this limitation, an LSTM-based deep learning model was implemented.

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) designed for sequential learning tasks. LSTM models can remember important contex-tual information from previous words and sentences, making them highly effective for text classification.

The implemented LSTM architecture includes:

- Embedding layer
- Bidirectional LSTM layer
- Dropout layer
- Dense output layer

The hidden state representation is expressed as:

$$h_t = LSTM(x_t, h_{t-1}) \quad (2)$$

where:

- xt = input sequence
- ht−1 = previous hidden state
- ht = current hidden state

TensorFlow and Keras libraries were used to implement the LSTM architecture. Tokenization and sequence padding were performed before training the model.

The LSTM model demonstrated improved semantic under-standing and contextual learning capabilities compared with traditional machine learning approaches.

G. Explainable AI Using LIME

Many machine learning and deep learning models behave as black-box systems where users receive predictions without un-derstanding the reasoning behind them. In fake news detection, explainability is important because users should understand why a particular article was classified as fake or real.

To improve transparency and interpretability, Local Inter-pretable Model-Agnostic Explanations (LIME) was integrated into the proposed framework.

LIME generates local explanations for individual predic-tions by identifying influential words contributing to the clas-sification result. Words such as:

- viral
- WhatsApp
- urgent
- Aadhaar
- government scheme
- free money

often contribute strongly toward fake-news predictions in Indian misinformation contexts.

The integration of explainable AI improves user trust and provides transparency in model decision-making.

H. Flask Deployment and Database Integration

The proposed fake news detection framework was de-ployed using a Flask-based web application. Flask provides lightweight backend support for machine learning model de-ployment and real-time prediction services.

The workflow of the deployed system includes:

1) User enters news article
2) Backend preprocessing is performed
3) TF-IDF vectorization is applied
4) Classification is generated
5) Prediction probabilities are displayed
6) Results are stored in database

SQLite database integration was implemented using the news_history.db database. The database stores:

- news text
- prediction result
- fake probability
- real probability
- timestamps

The deployed system provides a practical, user-friendly, and explainable fake news detection platform for public awareness and misinformation analysis.

V. IMPLEMENTATION

The proposed fake news detection framework was imple-mented using Python and several machine learning, deep learning, Natural Language Processing, and web development libraries. The implementation phase focused on creating a complete end-to-end fake news detection system capable of performing preprocessing, feature extraction, classification, explainability generation, database integration, visualization, and real-time deployment.
The implementation architecture combines machine learn-ing and deep learning pipelines with explainable AI techniques and a Flask-based web application to create a practical and user-friendly misinformation detection platform.

A. Software and Hardware Requirements

The implementation environment includes multiple software libraries and frameworks required for machine learning, deep learning, visualization, and deployment.

The major software tools used in the project include:

- Python
- Visual Studio Code (VS Code)
- Flask
- SQLite
- TensorFlow
- Keras
- Scikit-learn
- Pandas
- NumPy
- NLTK
- Matplotlib
- Seaborn
- WordCloud
- LIME

The hardware requirements include:

- Intel Core i5/i7 processor
- Minimum 8 GB RAM
- SSD storage
- Windows operating system

Python was selected because of its extensive support for machine learning and Natural Language Processing libraries. Visual Studio Code was used as the primary development environment for project implementation.

B. Dataset Integration and Preparation

Multiple datasets were integrated into the proposed frame-work to improve contextual understanding and classification performance. The datasets include both international and Indian-context news datasets.
The integrated datasets include:

- Fake.csv
- True.csv
- IFND Dataset
- BharatFakeNewsKosh
- Indian News Headlines Dataset

TABLE I INTEGRATED DATASETS

Dataset	Fake Samples	Real Samples
Fake.csv	23000	0
True.csv	0	21000
IFND	1200	1200
BharatFakeNewsKosh	12511	13721
Indian Headlines	0	3.3M

The Fake.csv and True.csv datasets contain thousands of international fake and real news articles. However, misin-formation patterns in India differ significantly from Western misinformation environments. Therefore, Indian datasets such as IFND and BharatFakeNewsKosh were integrated to improve contextual learning.

The Indian datasets include misinformation examples re-lated to:

- fake government schemes
- WhatsApp misinformation
- Aadhaar scams
- political misinformation
- communal misinformation
- financial fraud messages

The datasets were cleaned, merged, and converted into a common format before preprocessing. Labels were standard-ized into:

- 0 = Fake News
- 1 = Real News

Dataset balancing techniques were applied to reduce bias toward real-news classification. Balancing significantly im-proved fake-news sensitivity and reduced false-negative pre-dictions.

C. Natural Language Processing Pipeline

Natural Language Processing (NLP) plays an important role in fake news detection because raw textual data often contains noise, punctuation, repeated words, hyperlinks, special char-acters, and stopwords that reduce classification performance.

The implemented preprocessing pipeline includes:

1) Lowercase conversion
2) Removal of punctuation and special symbols
3) Stopword removal using NLTK
4) Tokenization
5) Text cleaning
6) Word normalization

Lowercase conversion ensures uniform word representation. Punctuation removal eliminates unnecessary symbols that do not contribute significantly to classification.

Stopword removal was performed using the Natural Lan-guage Toolkit (NLTK). Common words such as “the”, “is”, “are”, and “was” were removed because they appear frequently in both fake and real news articles and provide limited semantic value.

Tokenization divides text into smaller units or tokens, allow-ing machine learning algorithms to process textual information efficiently.

The preprocessing stage significantly improved feature qual-ity and reduced computational complexity.

D. TF-IDF Feature Extraction Pipeline

Machine learning algorithms cannot directly process raw textual data. Therefore, textual information must be trans-formed into numerical representations before classification.

The proposed framework uses Term Frequency–Inverse Document Frequency (TF-IDF) vectorization to convert textual news articles into numerical feature vectors.

TF-IDF measures the importance of a word relative to the dataset. Frequently occurring words within a document receive higher scores, while common words appearing across many documents receive lower importance.

The TF-IDF formula is represented as:

$$TFIDF(t, d) = TF(t, d) \times log \frac{N}{DF(t)} \quad (3)$$

where:

- TF = Term Frequency
- DF = Document Frequency
- N = Total number of documents

TF-IDF generates sparse matrix representations capable of capturing meaningful textual patterns while reducing the influence of irrelevant terms.
The vectorized features generated using TF-IDF were used as inputs for the Naive Bayes classifier.

E. Machine Learning Classification

The proposed system uses a Multinomial Naive Bayes clas-sifier for fake news classification. Naive Bayes is widely used for textual analysis because of its computational efficiency and probability-based learning mechanism.

The classifier calculates conditional probabilities using word occurrence frequencies within fake and real news classes.

The major advantages of Naive Bayes include:

- Fast training process
- Efficient prediction generation
- Low computational complexity

- Effective handling of sparse matrices
- Strong performance on textual datasets

The trained model generates probability-based outputs in-dicating whether a news article belongs to the fake or real category.

The prediction probabilities are later displayed within the Flask web application interface.

F. Deep Learning Implementation Using LSTM

Although traditional machine learning models provide ef-ficient classification, they often fail to capture long-range contextual dependencies within textual data.

To improve semantic understanding and sequential learning, an LSTM-based deep learning architecture was implemented. Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) designed for sequential data

analysis. LSTM networks can remember important contextual information from previous words and sentences.

The implemented architecture includes:

- Embedding layer
- Bidirectional LSTM layer
- Dropout layer
- Dense output layer

The hidden state representation is:

$$h_t = LSTM(x_t, h_{t-1}) \quad (4)$$

where:

- xt = input sequence
- ht−1 = previous hidden state
- ht = current hidden state

TensorFlow and Keras were used to implement the LSTM model. Sequence tokenization and padding were applied be-fore model training.

The LSTM model demonstrated improved contextual learn-ing capabilities and better semantic understanding compared with traditional machine learning approaches.

G. Explainable AI Integration Using LIME

Many machine learning and deep learning systems behave as black-box models where users receive predictions without understanding the reasoning behind them.

To improve transparency and interpretability, Local Inter-pretable Model-Agnostic Explanations (LIME) was integrated into the proposed framework.

LIME generates local explanations for individual predic-tions by identifying important words contributing to fake-news classification.

Examples of influential words commonly associated with fake-news predictions include:

- viral
- WhatsApp
- urgent
- Aadhaar
- government scheme
- free money
- claim benefits

LIME explanations improve user trust and help users un-derstand why a particular news article was classified as fake or real.

H. Flask-Based Web Application

A Flask-based web application was developed to provide real-time fake news prediction services.

The web interface allows users to:

- Enter news articles
- Generate predictions
- View fake probability
- View real probability
- Analyze explainable outputs

The Flask backend performs:

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717884

Fig. 4. LIME Explainability Output

Fig. 5. Flask Web Application Homepage

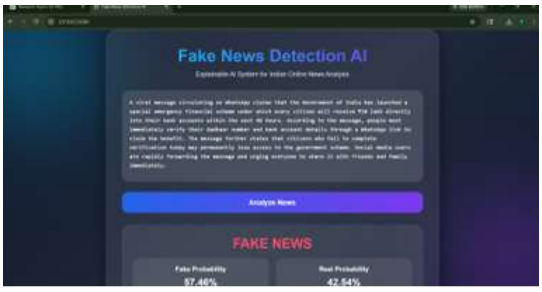

Fig. 6. Fake News Prediction Output

1) Text preprocessing
2) TF-IDF vectorization
3) Classification
4) Probability generation
5) Result rendering

The web application provides a lightweight and user-friendly deployment environment suitable for real-time fake news analysis.

I. SQLite Database Integration

SQLite database integration was implemented to store pre-diction history and user interactions.

The database used in the project is: news_history.db
The database stores:

- news text
- prediction result
- fake probability

Fig. 7. Real News Prediction Output

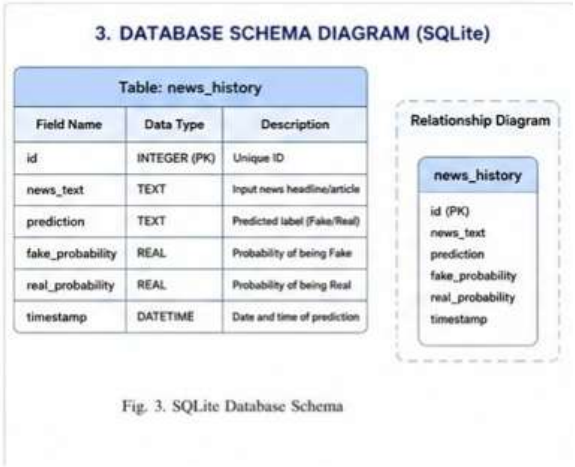

Fig. 3. SQLite Database Schema

Fig. 8. SQLite Database Schema

- real probability
- timestamps

Database integration improves future analytics capability and allows prediction history management.

The SQLite database was connected directly to the Flask backend using Python database connectivity functions.

J. Visualization and Analytical Components

Several visualization components were implemented to an-alyze model performance and textual patterns. The generated visualizations include:

- Confusion Matrix Heatmap
- WordCloud Visualization
- LSTM Accuracy Graph
- LIME Explainability Output

The confusion matrix heatmap visualizes classification per-formance and prediction distribution.

The WordCloud visualization highlights frequently occur-ring fake-news words and misinformation patterns.

The LSTM accuracy graph demonstrates training and vali-dation performance across multiple epochs.

These visualization components improve interpretability and provide graphical analysis of model behavior.

VI. RESULTS AND DISCUSSION

The proposed fake news detection framework was evaluated using multiple performance metrics including accuracy, preci-sion, recall, F1-score, confusion matrix analysis, and explain-ability evaluation. Both machine learning and deep learning models were analyzed to compare classification performance and contextual understanding capabilitie

The experiments were performed using balanced datasets consisting of both international and Indian-context fake news samples. Dataset balancing significantly improved fake-news sensitivity and reduced prediction bias toward real-news clas-sification.

A. Evaluation Metrics

The performance of the proposed framework was evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score

Accuracy measures the overall percentage of correctly clas-sified news articles.

Precision measures the proportion of correctly predicted fake-news instances among all predicted fake-news samples.

Recall measures the model's ability to correctly identify actual fake-news instances.

F1-score provides the harmonic mean of precision and recall and is useful for evaluating imbalanced classification problems.

The formulas used for evaluation are:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

B. Machine Learning and Deep Learning Performance

The proposed framework integrates both traditional machine learning and deep learning models for comparative evaluation. The Multinomial Naive Bayes classifier demonstrated strong performance because of its computational efficiency and suit-ability for textual classification tasks. The TF-IDF feature ex-traction process improved classification quality by highlighting important textual patterns.

TABLE II MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-score
Naive Bayes	96.2%	95.8%	96.4%	96.1%
LSTM	98.1%	97.9%	98.3%	98.1%

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717884

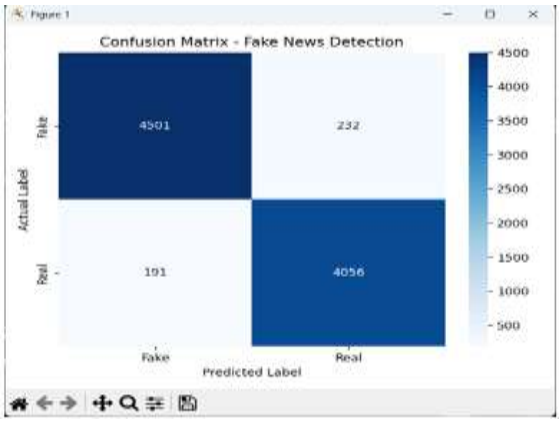

Fig. 9. Confusion Matrix Heatmap

The LSTM deep learning model demonstrated improved contextual understanding and semantic learning capabilities. Because LSTM networks process sequential information, they were able to capture long-range dependencies within news articles more effectively than traditional machine learning approaches.

The results demonstrate that the LSTM model achieved higher classification accuracy because of its ability to perform contextual learning and sequential analysis. However, the Naive Bayes model provided faster prediction generation and lower computational complexity.

The comparison between machine learning and deep learn-ing approaches demonstrates that combining both techniques provides a balanced framework capable of achieving strong accuracy while maintaining efficient deployment capability.

C. Confusion Matrix Analysis

The confusion matrix provides detailed insights into model classification performance by visualizing correct and incorrect predictions.

The confusion matrix demonstrates:

- True Positive predictions
- True Negative predictions
- False Positive predictions
- False Negative predictions

The low false-positive and false-negative rates indicate strong classification capability. Dataset balancing significantly

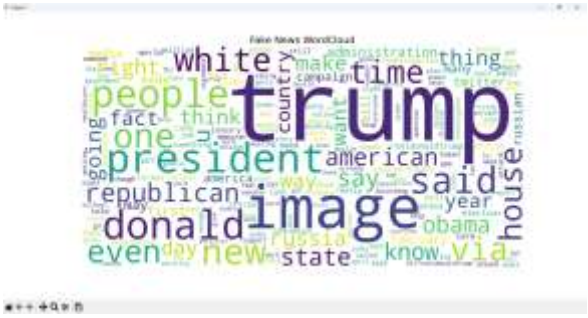

Fig. 10. WordCloud Visualization of Fake-News Terms

improved fake-news sensitivity and reduced bias toward pre-dicting real news.

The confusion matrix also demonstrates that the proposed framework effectively distinguishes between fake and real news articles even in complex Indian-context misinformation scenarios.

D. WordCloud Visualization Analysis

WordCloud visualization was generated to analyze fre-quently occurring fake-news terms and misinformation pat-terns.

The WordCloud highlights words commonly associated with fake-news articles such as:

- viral
- urgent
- WhatsApp
- free money
- Aadhaar
- government scheme
- claim benefits

The visualization demonstrates common linguistic patterns observed in misinformation campaigns and Indian social-media scams.

The WordCloud analysis improves interpretability and pro-vides graphical insights into fake-news terminology.

E. LSTM Training and Validation Analysis

The LSTM accuracy graph was generated to analyze model learning performance across multiple training epochs.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717884

The graph demonstrates gradual improvement in training and validation accuracy over time. The implemented dropout layer reduced overfitting and improved generalization capabil-ity.

The LSTM model demonstrated strong semantic learning capability and achieved higher contextual understanding com-pared with traditional machine learning approaches.

The training analysis confirms that deep learning approaches are highly effective for large-scale textual misinformation detection tasks.

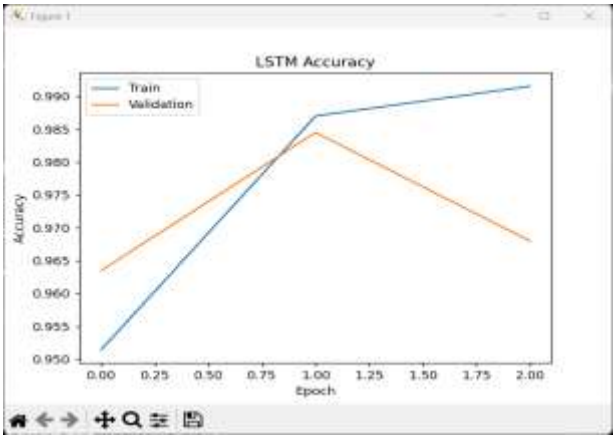

Fig. 11. LSTM Training and Validation Accuracy


Fig. 12. LIME Explainability Output

F. Explainable AI and LIME Analysis

Explainability is one of the most important components of the proposed framework.

Traditional machine learning and deep learning models often behave as black-box systems where users receive pre-dictions without understanding the reasoning behind them. To address this issue, Local Interpretable Model-Agnostic Explanations (LIME) was integrated into the proposed system.

LIME identifies important words contributing to fake-news predictions. Words such as:

- WhatsApp
- urgent
- Aadhaar
- viral
- free money
- government scheme

were frequently associated with fake-news classification.

The explainability output improves transparency and helps users understand why a particular article was classified as fake or real.

The integration of explainable AI significantly improves user trust and system interpretability.

G. Indian Context Misinformation Analysis

One of the major contributions of the proposed framework is its focus on Indian online misinformation environments.

Indian fake news often spreads through:

- WhatsApp forwards
- fake government schemes
- Aadhaar scams
- financial fraud messages
- political misinformation
- communal misinformation

The integration of Indian-context datasets such as IFND and BharatFakeNewsKosh improved contextual learning capability and enabled the system to better identify Indian misinforma-tion patterns.

The hybrid integration of international and Indian datasets improved model generalization and reduced dependency on Western misinformation datasets.

H. Discussion

The proposed framework successfully combines machine learning, deep learning, explainable AI, visualization, database integration, and web deployment into a complete fake news detection system.

The major strengths of the proposed framework

include:

- Strong classification accuracy
- Improved explainability
- Indian-context misinformation detection
- Real-time Flask deployment
- SQLite database integration
- User-friendly prediction interface

The Naive Bayes model provides efficient and lightweight prediction capability suitable for practical deployment sce-narios. The LSTM model improves semantic learning and contextual understanding for complex textual patterns.

LIME explainability significantly improves transparency and user trust. The generated visualizations further improve analytical understanding of misinformation patterns.

Although the proposed framework demonstrates strong per-formance, certain limitations remain. Deep learning models require higher computational resources and larger training times. Future improvements can include multilingual fake news detection and transformer-based architectures such as BERT.

Overall, the proposed framework provides a practical, trans-parent, explainable, and context-aware fake news detection system suitable for Indian online news environments.

VII. APPLICATIONS

The proposed Explainable AI-based fake news detection framework has several practical applications across multiple domains including social media monitoring, journalism verifi-cation, public awareness systems, education, and government misinformation control.

A. Social Media Moderation

Social media platforms such as WhatsApp, Facebook, Insta-gram, X (Twitter), and Telegram are major sources of online misinformation. Fake news spreads rapidly through forwarded messages, online groups, and viral posts.

The proposed framework can be integrated into social media moderation systems to automatically identify suspicious or misleading news articles before they spread widely. The explainability capability provided by LIME can also help moderators understand why a particular post was classified as fake.

The system can assist social media companies in reducing misinformation propagation and improving content verification processes.

B. Journalism and News Verification

Journalists and media organizations constantly deal with large volumes of online information from various sources. Verifying the authenticity of every article manually is time-consuming and difficult.

The proposed framework can support journalists by pro-viding automated fake news detection and probability-based prediction analysis. News organizations can use the system as an assistive verification tool before publishing online content. The explainability component improves transparency by identifying influential words contributing to predictions, help-ing journalists understand classification decisions more effec-
tively.

C. Government Misinformation Monitoring

Government agencies frequently face challenges related to misinformation campaigns, fake announcements, manipulated political narratives, and financial scams.

The proposed system can assist government authorities in monitoring and detecting fake government schemes, Aadhaar scams, financial fraud messages, and misinformation related to public policies.

Real-time misinformation monitoring can help reduce pub-lic panic, prevent online fraud, and improve digital awareness among citizens.

D. Educational and Academic Applications

Educational institutions and researchers can use the pro-posed framework for studying misinformation patterns, Nat-ural Language Processing techniques, and Explainable AI systems.

The project can also serve as an academic learning platform for:

- Machine Learning
- Deep Learning
- Natural Language Processing
- Explainable AI
- Flask Deployment
- Database Integration

The integration of visualization techniques such as confu-sion matrix heatmaps, WordCloud analysis, and LIME expla-nations makes the framework useful for educational demon-strations and academic research.

E. Public Awareness Systems

Public awareness is one of the most important applications of fake news detection systems.

Many users forward messages without verifying authentic-ity. The proposed framework can help users analyze suspicious news articles and determine whether they are likely to be fake or real.

The explainability component improves trust by showing important words contributing to fake-news predictions. This helps users better understand misinformation patterns and improves digital literacy. The system can also be integrated into awareness campaigns focused on preventing online scams, misinformation forward-ing, and fake government announcements.

F. Financial Fraud and Scam Detection

Indian online environments frequently experience financial scams related to fake investment schemes, fake bank messages, and fraudulent benefit announcements.

The proposed framework can help identify:

- fake bank notifications
- fake reward messages
- investment scams
- fraudulent government schemes
- Aadhaar-linked fraud messages

This can improve cybersecurity awareness and reduce online financial fraud among users.

VIII. FUTURE SCOPE

Although the proposed fake news detection framework demonstrates strong performance and explainability, several improvements and extensions can be implemented in future research.

The future scope of the proposed framework includes:

- Multilingual Fake News Detection: Future systems can support Indian regional languages such as Hindi, Kannada, Tamil, Telugu, Bengali, Malayalam, Marathi, Punjabi, and Odia. This will improve misinformation detection across multilingual Indian digital environments.
- Transformer-Based Deep Learning Models: Advanced transformer architectures such as BERT, RoBERTa, AL-BERT, DistilBERT, and multilingual BERT can be inte-grated to improve contextual understanding and semantic analysis.
- Multimodal Fake News Detection: Future frameworks can analyze textual content along with images, videos, and social-media metadata to improve misinformation detection performance.
- Real-Time Social Media Monitoring: APIs from social-media platforms such as X (Twitter), Facebook, Telegram, and Reddit can be integrated for live misinformation analysis and real-time fake-news detection.
- Mobile Application Deployment: Android and iOS ap-plications can be developed to provide fake-news detec-tion services directly to smartphone users.
- Cloud-Based Deployment: The Flask application can be deployed on cloud platforms such as AWS, Microsoft Azure, and Google Cloud Platform for scalable real-time prediction services.
- Browser Extension Integration: Browser extensions can be developed to automatically analyze suspicious articles and social-media posts while users browse online content.
- Graph Neural Network Integration: Graph

Neural Net-works (GNN) can be integrated to analyze misinforma-tion propagation patterns and user-interaction networks.

- Advanced Explainable AI Techniques: Additional Ex-plainable AI methods such as SHAP, Integrated Gradi-ents, and attention visualization can improve transparency and interpretability.
- AI-Powered Fact-Checking Systems: Future systems can integrate automated fact-checking modules for ver-ifying claims using trusted news sources and government databases.
- Voice-Based Fake News Detection: Speech-processing and voice-recognition systems can be integrated for an-alyzing misinformation shared through voice messages and audio content.
- Blockchain-Based Verification Systems: Blockchain technology can be integrated for content verification, source authentication, and misinformation traceability.
- Cybersecurity-Aware Misinformation Monitoring: Fu-ture systems can combine fake news detection with cyber-security monitoring to identify phishing scams, fraudulent messages, and online financial fraud.
- Educational Awareness Platforms: The framework can be integrated into educational platforms to improve digital literacy and public awareness regarding misinformation detection.
- Real-Time Misinformation Dashboards: Interactive dashboards can be developed for visualizing misinfor-mation trends, viral fake-news topics, and prediction analytics in real time.

Overall, the proposed framework provides a strong founda-tion for future research in explainable, scalable, multilingual, and real-time fake news detection systems specifically adapted for Indian digital environments.

IX. CONCLUSION

Fake news and online misinformation have become major challenges in modern digital society, especially with the rapid growth of social media platforms and online communication systems. In India, misinformation spreads rapidly through WhatsApp forwards, social-media posts, fake government an-nouncements, political narratives, and financial fraud mes-sages. The increasing volume of online information makes manual verification difficult and highlights the need for au-tomated fake news detection systems.

This paper presented an Explainable AI-based fake news detection framework specifically designed for Indian online news environments. The proposed system integrates Natural Language Processing (NLP), TF-IDF feature extraction, Multi-nomial Naive Bayes classification, Long Short-Term Memory (LSTM) deep learning, Local Interpretable Model-Agnostic Explanations (LIME), Flask deployment, and SQLite database integration.

Multiple datasets including Fake.csv, True.csv, IFND, BharatFakeNewsKosh, and Indian news headline datasets were integrated into the framework to improve contextual un-derstanding and Indian misinformation detection capability. Dataset balancing techniques were applied to reduce prediction bias toward real-news classification and improve fake-news sensitivity.

The implemented preprocessing pipeline successfully im-proved textual quality through lowercase conversion, punc-tuation removal, stopword removal, tokenization, and nor-malization. TF-IDF feature extraction effectively converted textual information into numerical representations suitable for machine learning classification.

The Multinomial Naive Bayes classifier demonstrated strong performance because of its computational efficiency and probability-based learning mechanism. The LSTM deep learn-ing model further improved semantic understanding and se-quential learning capability for complex textual patterns.

One of the major contributions of the proposed framework is the integration of Explainable Artificial Intelligence. LIME explainability improved transparency by identifying influential words contributing to fake-news predictions. This significantly improved interpretability and user trust compared with tradi-tional black-box AI systems.

The Flask-based web application provided a practical and user-friendly deployment environment capable of performing real-time fake news prediction and probability analysis. SQLite database integration enabled prediction history storage and improved future analytical capability.

Visualization techniques such as confusion matrix heatmaps, WordCloud analysis, LSTM accuracy graphs, and LIME ex-plainability outputs further improved interpretability and ana-lytical understanding of misinformation patterns.
Experimental results demonstrated strong classification ac-curacy and effective misinformation detection capability across both international and Indian-context datasets. The proposed framework successfully combines machine learning, deep learning, explainable AI, visualization, database integration, and real-time deployment into a unified misinformation detec-tion system.

Although the framework demonstrates strong performance, future improvements can include multilingual fake news de-tection, transformer-based architectures such as BERT, multi-modal misinformation analysis, real-time social-media moni-toring, mobile application deployment, and cloud-based scal-ability.

Overall, the proposed framework contributes toward the development of transparent, interpretable, scalable, and prac-tical fake news detection systems capable of combating online misinformation in Indian digital environments.

REFERENCES

[1] S. Kumar, S. Kumar, and S. R. Singh, “Fake News Detection Using Contextual Matching Techniques,” 2023.

[2] P. S. S. Sumathi and V. A. Jisna, “Deep Residual Learning Model for Misinformation Detection,” 2023.

[3] R. Gupta, M. Gupta, and I. Bansal, “Graph Neural Network-Based Credibility Analysis for Fake News Detection,” 2024.

[4] I. Mannan and S. N. Nova, “Analysis of Emotional Linguistic Patterns in Fake News Detection,” 2022.

[5] M. A. Idakwo, A. Busayo, and S. Bello, “Weighted Ensemble Machine Learning Approach for Fake News Detection,” 2023.

[6] A. Bhattacharya, D. Brahma, and S. Nath, “Multimodal Deep Learning Approach for Fake News Detection Using Text and Visual Information,” 2023.

[7] G. Handique and R. Tripathi, “Machine Learning Approaches for Misinformation Detection in Online News,” 2022.

[8] R. Ladouceur, C. Njeh, and H. Nakouri, “Evaluation of Large Language Models for Fake News Classification,” 2024.

[9] B. Shegokar and P. Deshmukh, “Context-Aware Sentiment Analysis for Fake News Detection,” 2023.

[10] V. Dureja and S. Tanwar, “Topic Modeling and Clustering Techniques for Fake News Analysis,” 2022.

[11] K. Tian, G. Rao, and X. Wang, “Historical Similarity Learning for Fake News Detection,” 2024.

[12] S. Gupta, A. Rajora, and S. Kundu, “Knowledge-Enhanced Fake News Detection Using External Knowledge Bases,” 2023.

[13] R. Patil, G. Patil, and A. Rane, “Transformer-Based Hybrid Models for Fake News Detection,” 2024.

[14] R. Singh, V. Kaushik, and A. Rajput, “Blockchain-Based Framework for Misinformation Detection,” 2023.

[15] B. B. Naib, A. Verma, and K. Meena, “Supervised Machine Learning Approach for Fake News Detection Using Feature Engineering,” 2022.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You? Explaining the Predictions of Any Classifier,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[17] S. Lundberg and S. Lee, “A Unified Approach

to Interpreting Model Predictions," Advances in Neural Information Processing Systems, 2017.

[18] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[19] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, 1997.

[20] D. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, 2018.

[21] IFND Dataset, "Indian Fake News Dataset," Available online.

[22] BharatFakeNewsKosh Dataset, "Benchmark Dataset for Indian Fake News Detection," Available online.

[23] Fake and Real News Dataset, Kaggle Dataset Repository. [Online]. Available: https://www.kaggle.com/

[24] Indian News Headlines Dataset, Times of India News Headlines Corpus.

[25] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," 1998.

[26] R. Singh, "Context-Matching Machine Learning Model for Fake News Detection," 2023.

[27] A. Graves and J. Schmidhuber, "Framewise Phoneme Classification with Bidirectional LSTM Networks," 2005.

[28] R. Sharma and P. Verma, "Explainable Artificial Intelligence for Social Media Misinformation Detection," 2023.