IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

A Web-Based Intelligent Framework for Automated Fake News Detection Using Logistic Regression and NLP Pipeline

Mirza Mohammed Amin Mohammed Raza
Student, Master of Computer Applications (MCA)
University of Mumbai, Mumbai, Maharashtra, India

Abstract: *In this research study, we developed a web-based automated framework integrated with a Natural Language Processing (NLP) pipeline for efficient fake news detection. Traditional text classification models often face challenges in accurately identifying misinformation due to evolving linguistic patterns and large, unstructured textual datasets. By incorporating a Term Frequency-Inverse Document Frequency (TF-IDF) vectorization pipeline, our model enhances text feature representation and lexical weight extraction, while a regular-expression-based cleaning script captures structural context for precise text classification.*

We trained and evaluated a Logistic Regression classifier on a consolidated dataset consisting of thousands of public articles categorized into real and fake news classes. Experimental results demonstrate that the proposed approach yields high performance, achieving an overall classification accuracy of 98.6% alongside superior metrics across precision, recall, and F1-score while significantly reducing false positives. The optimized backend pipeline was seamlessly integrated into a live, interactive web dashboard using the Streamlit framework, incorporating real-time performance evaluation and live reference news verification through API connections. This work highlights the high utility of pairing lightweight linear classifiers with robust NLP preprocessing pipelines to address modern digital misinformation challenges at scale.

Keywords: Fake news detection, Logistic Regression, Natural Language Processing (NLP), TF-IDF Vectorization, Streamlit web application.

I. INTRODUCTION

In today's digital era, online news platforms and social media have become primary sources of information for millions of people. While these platforms enable rapid dissemination of information, they also facilitate the spread of fake news, misinformation, and misleading content. Fake news can influence public opinion, create social unrest, manipulate political processes, and undermine trust in legitimate media sources. As the volume of digital content continues to grow, manually verifying the authenticity of news articles becomes increasingly difficult. Therefore, automated fake news detection systems have emerged as an important research area in the fields of Natural Language Processing (NLP) and Machine Learning.

Traditional fact-checking methods rely heavily on human experts and are often time-consuming and resource-intensive. Machine learning techniques provide a scalable alternative by analyzing textual patterns and classifying news articles as real or fake. However, developing an effective fake news detection system remains challenging due to the diversity of writing styles, evolving misinformation strategies, and the large volume of online content generated daily.

To address these challenges, this research proposes a Fake News Detection System based on Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction and Logistic Regression classification. The system analyzes news article titles and content, transforms textual information into numerical representations, and classifies news articles

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

based on learned patterns from historical data. Additionally, a web-based dashboard developed using Streamlit enables real-time news analysis and integrates live news retrieval through NewsAPI, providing users with an interactive platform for detecting potentially misleading information.

A. Problem Statement

The rapid growth of online news sources and social media platforms has significantly increased the spread of fake and misleading information. Existing manual verification methods are often slow, labor-intensive, and unable to keep pace with the volume of news generated every day. Furthermore, users frequently encounter news from unverified sources, making it difficult to determine credibility and authenticity.

The primary problem addressed in this research is the development of an accurate, efficient, and scalable fake news detection system capable of automatically classifying news articles as real or fake. The system must effectively process textual information, identify linguistic patterns associated with misinformation, and provide reliable predictions in real-time scenarios. This research aims to leverage machine learning techniques to reduce misinformation exposure and support informed decision-making among users.

B. Significance of Research

This research contributes to the growing field of automated misinformation detection by developing a machine learning-based framework for fake news classification. The proposed system utilizes TF-IDF vectorization and Logistic Regression to analyze textual content and identify patterns associated with fake and real news articles.

The significance of this research lies in its ability to provide a practical solution for detecting misinformation efficiently. By integrating a real-time web application and live news retrieval functionality, the system extends beyond theoretical research and demonstrates practical usability. The findings of this study may also serve as a foundation for future research involving advanced machine learning models, deep learning approaches, and multilingual fake news detection systems.

C. Motivation of Research

The increasing prevalence of fake news across digital platforms has become a major societal concern. False information can influence public perception, spread panic, affect elections, damage reputations, and create confusion during critical events. These consequences highlight the urgent need for reliable automated systems capable of identifying misleading information before it reaches a wider audience.

The motivation behind this research stems from the desire to develop a practical machine learning solution that can assist users in evaluating the credibility of news content. The availability of large-scale news datasets and advances in Natural Language Processing techniques provide an opportunity to build effective detection models. Additionally, the integration of real-time news analysis through a web-based interface enhances the usability and applicability of the proposed system.

Another motivating factor is the growing importance of artificial intelligence in combating misinformation. By applying machine learning techniques to fake news detection, this research demonstrates how technology can contribute to improving information reliability and promoting digital literacy.

D. Scope of Research

This research focuses on the development of a fake news detection system using machine learning techniques. The system utilizes news article titles and textual content as input data and employs TF-IDF vectorization for feature extraction and Logistic Regression for classification.

The scope of this research includes dataset preprocessing, feature extraction, model training, performance evaluation, model deployment, and real-time news analysis through a Streamlit-based dashboard. The system is designed to classify English-language news articles as either real or fake and provide probability-based predictions to users.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

E. Limitations

Despite the effectiveness of the proposed system, several limitations exist. The model relies primarily on textual content and does not verify factual claims through external knowledge bases or fact-checking websites. Therefore, highly sophisticated misinformation that closely resembles legitimate news may still pose classification challenges.

The system is trained using a specific dataset consisting of real and fake news articles, which may limit its ability to generalize across all domains, languages, or emerging misinformation patterns. Furthermore, the performance of the model depends on the quality and diversity of the training data. Since Logistic Regression is a traditional machine learning algorithm, more advanced deep learning models may achieve improved performance on complex datasets.

Research Objectives

1. To develop a machine learning-based fake news detection system using Logistic Regression.
2. To preprocess and transform textual news data using TF-IDF vectorization.
3. To evaluate model performance using accuracy, precision, recall, and F1-score metrics.
4. To implement a real-time fake news detection dashboard using Streamlit.
5. To integrate live news retrieval and analysis through NewsAPI.

Research Questions

1. Can Logistic Regression effectively classify news articles as real or fake using textual features?
2. How does TF-IDF feature extraction contribute to fake news detection performance?
3. What level of accuracy can be achieved using the proposed machine learning framework?
4. Can the developed system provide reliable predictions in real-time environments?
5. How can the integration of live news retrieval improve the usability of fake news detection systems?

II. LITERATURE REVIEW

Fake news detection has emerged as a significant research area due to the rapid growth of online news platforms and social media. Researchers have explored various machine learning and natural language processing techniques to automatically identify misleading and fabricated information. Early approaches primarily relied on traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), Decision Trees, and Logistic Regression. These methods demonstrated promising results when combined with effective text preprocessing and feature extraction techniques.

Feature extraction plays a crucial role in text classification tasks. Among various approaches, Term Frequency–Inverse Document Frequency (TF-IDF) has been widely adopted because of its ability to represent textual data numerically while preserving the importance of meaningful words within documents. Studies have shown that TF-IDF effectively transforms unstructured text into feature vectors suitable for machine learning models, improving classification performance in fake news detection systems.

Recent research has also explored deep learning approaches such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks. These models can automatically learn complex textual patterns and contextual relationships within news articles. Although deep learning methods often achieve high accuracy, they require large datasets, extensive computational resources, and longer training times, making them less practical for lightweight and real-time applications.

Logistic Regression remains one of the most widely used classification algorithms for text-based prediction tasks because of its simplicity, efficiency, and strong performance on high-dimensional datasets. When combined with TF-IDF vectorization, Logistic Regression has demonstrated competitive accuracy while maintaining lower computational complexity compared to deep learning approaches. This makes it particularly suitable for practical fake news detection systems where fast prediction and scalability are important.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

In addition to classification models, recent studies have emphasized the importance of real-time misinformation detection. Researchers have explored integrating machine learning models with web-based applications and news aggregation platforms to provide instant credibility assessment of online content. Such systems improve accessibility and allow users to analyze news articles directly through interactive interfaces.

Limitations of Previous Work

Despite significant advancements in fake news detection, several limitations remain. Traditional machine learning models often depend heavily on textual features and may struggle to identify highly sophisticated misinformation that closely resembles legitimate news content. Deep learning approaches, while powerful, require substantial computational resources, longer training times, and larger datasets to achieve optimal performance.

Many existing studies focus primarily on offline experimentation and benchmark evaluation rather than practical deployment. As a result, issues related to real-time prediction, user interaction, and system scalability remain insufficiently explored. Furthermore, several models are trained on limited datasets that may not adequately represent evolving misinformation patterns found in real-world environments.

Another limitation is that many fake news detection systems focus solely on classification accuracy without considering usability and accessibility for end users. Consequently, there is a growing need for lightweight, interpretable, and deployable solutions that can provide real-time analysis while maintaining satisfactory performance.

Gaps in Previous Studies

Although numerous machine learning and deep learning approaches have been proposed for fake news detection, several research gaps still exist. Many studies emphasize model performance but do not address practical implementation and real-time deployment. Additionally, the integration of machine learning models with live news retrieval systems remains relatively underexplored.

Existing research often prioritizes complex deep learning architectures that require significant computational resources, making them difficult to deploy in resource-constrained environments. There is also limited focus on developing lightweight models that balance prediction accuracy with computational efficiency. Furthermore, many systems do not provide confidence-based predictions or user-friendly visualization tools that can assist users in interpreting results.

These gaps highlight the need for a practical fake news detection framework that combines effective text processing, reliable classification, and real-time usability. The present research addresses these challenges through the integration of TF-IDF feature extraction, Logistic Regression classification, and a Streamlit-based interactive dashboard.

Datasets Used in Previous Studies

Various datasets have been used in fake news detection research. Popular datasets include the Fake and Real News Dataset, LIAR Dataset, FakeNewsNet, and ISOT Fake News Dataset. These datasets contain labeled news articles categorized as real or fake and have been widely utilized for training and evaluating machine learning models.

The Fake and Real News Dataset is particularly popular because it contains a substantial collection of authentic and fabricated news articles from diverse domains. The dataset provides textual information such as news titles and article content, making it suitable for Natural Language Processing and machine learning applications. In this research, the Fake.csv and True.csv datasets are utilized to develop and evaluate the proposed fake news detection model.

Feature Extraction Techniques

Feature extraction is a critical step in transforming textual information into numerical representations suitable for machine learning algorithms. Traditional methods include Bag-of-Words (BoW), N-grams, and TF-IDF representations. Among these techniques, TF-IDF is widely recognized for its ability to emphasize informative words while reducing the influence of commonly occurring terms.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Unlike simple frequency-based approaches, TF-IDF assigns weights based on both word frequency within a document and rarity across the entire dataset. This enables the model to focus on distinguishing terms that contribute significantly to classification performance. Due to its effectiveness and computational efficiency, TF-IDF has been adopted as the primary feature extraction technique in this research.

Summary of Existing Theories and Models

Over the years, fake news detection has evolved from traditional machine learning approaches to more advanced deep learning techniques. Classical algorithms such as Naïve Bayes, Support Vector Machines, Decision Trees, and Logistic Regression established the foundation for automated text classification. Later, deep learning architectures such as CNNs, RNNs, and LSTMs improved the ability to capture contextual and sequential information from textual data.

Despite these advancements, lightweight machine learning models remain attractive because they offer faster training, lower computational cost, and easier deployment. Logistic Regression, particularly when combined with TF-IDF vectorization, continues to demonstrate strong performance in text classification tasks. Building upon these foundations, the proposed research develops a fake news detection system that leverages TF-IDF feature extraction and Logistic Regression classification while providing real-time analysis through an interactive Streamlit dashboard and live news integration. This approach bridges the gap between research-oriented models and practical applications by offering an efficient, scalable, and user-friendly solution for combating online misinformation.

III. PROBLEM DEFINITION

The internet has transformed the way people access and consume information. News articles, blogs, online forums, and social media platforms provide instant access to information from around the world. While this rapid exchange of information has many advantages, it has also led to the widespread dissemination of fake news, misinformation, and misleading content. Fake news refers to deliberately fabricated or manipulated information presented as legitimate news with the intention of influencing public opinion, generating confusion, or achieving political, social, or financial objectives.

The growing prevalence of fake news has become a serious concern in modern society. False information can spread much faster than verified information, especially through social media platforms where users can share content instantly with large audiences. During critical events such as elections, public health emergencies, natural disasters, and social movements, fake news can create panic, mislead citizens, influence decision-making, and undermine trust in reliable news sources. The impact of misinformation extends beyond individuals and can affect governments, organizations, businesses, and entire communities.

Traditionally, the verification of news articles has relied on journalists, fact-checking organizations, and domain experts. Although these methods are effective, they are often time-consuming, labor-intensive, and unable to keep pace with the enormous volume of information generated every day. Millions of news articles, social media posts, and online reports are published daily, making manual verification increasingly impractical. As a result, there is a growing need for automated systems capable of detecting fake news efficiently and accurately.

Developing an automated fake news detection system presents several challenges. Fake news creators continuously modify their writing styles, vocabulary, and presentation techniques to make false information appear credible. Many fake articles closely resemble authentic news reports, making it difficult to distinguish between genuine and fabricated content. Furthermore, the language used in news articles is often complex and context-dependent, requiring sophisticated text analysis techniques to identify deceptive patterns.

Existing approaches for fake news detection include rule-based systems, traditional machine learning algorithms, and deep learning models. Rule-based systems rely on predefined patterns and keywords, but they often fail to adapt to evolving misinformation strategies. Deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have demonstrated strong performance; however, they typically require large

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

training datasets, extensive computational resources, and significant processing time. These requirements can limit their practicality for real-time applications and deployment in resource-constrained environments.

Another major challenge is transforming unstructured textual information into a format that machine learning algorithms can process effectively. News articles consist of large amounts of textual data containing different writing styles, sentence structures, and vocabulary. Therefore, effective feature extraction techniques are essential for capturing meaningful information while reducing noise and irrelevant content. Without proper feature representation, classification models may fail to achieve satisfactory performance.

The problem addressed in this research is the design and implementation of an automated fake news detection system that can accurately classify news articles as real or fake using machine learning techniques. The system should be capable of processing textual data efficiently, extracting informative features, identifying patterns associated with misinformation, and providing reliable predictions with minimal computational overhead. In addition, the system should support real-time analysis and offer an intuitive interface that enables users to evaluate news content quickly and effectively.

To address these challenges, this research proposes a machine learning-based approach that combines Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction with Logistic Regression classification. TF-IDF is used to convert textual content into numerical feature vectors by measuring the importance of words within news articles, while Logistic Regression serves as the classification model for distinguishing between real and fake news. The model is trained using labeled datasets containing authentic and fabricated news articles, enabling it to learn patterns associated with each category.

Furthermore, the proposed system incorporates a Streamlit-based web application that allows users to input news content and receive instant classification results. The system also integrates live news retrieval through NewsAPI, enabling real-time analysis of current news articles. By combining machine learning, natural language processing, and real-time deployment, the proposed solution aims to provide an effective, scalable, and user-friendly approach for combating the growing problem of online misinformation.

Ultimately, this research seeks to contribute to the development of intelligent systems that support information verification, enhance digital literacy, and help users make informed decisions when consuming online news. By reducing reliance on manual fact-checking and enabling rapid identification of potentially misleading content, the proposed fake news detection system can play an important role in promoting a more trustworthy and reliable digital information ecosystem.

IV. OBJECTIVES OF THE RESEARCH

The primary objective of this research is to develop an efficient and reliable fake news detection system using machine learning techniques. The proposed system aims to automatically classify news articles as real or fake by analyzing their textual content and identifying patterns associated with misinformation.

The specific objectives of this research are:

1. To develop a machine learning-based fake news detection model using Logistic Regression for binary classification of news articles.
2. To preprocess and clean textual news data to improve the quality and consistency of the dataset.
3. To apply Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction for converting textual news content into numerical representations suitable for machine learning algorithms.
4. To evaluate the performance of the proposed model using standard classification metrics such as Accuracy, Precision, Recall, and F1-Score.
5. To develop a user-friendly web application using Streamlit for real-time fake news analysis.
6. To integrate live news retrieval through NewsAPI to enhance the practicality and usability of the system.
7. To provide an automated solution that assists users in identifying potentially misleading or false information and promotes informed decision-making.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Scope of the Research

This research focuses on the design, implementation, and evaluation of a machine learning-based fake news detection system. The system is designed to analyze textual news content and classify it as either real or fake using supervised machine learning techniques.

The scope of this study includes:

1. Collection and utilization of labeled fake and real news datasets for model training and testing.
2. Data preprocessing techniques such as handling missing values, text cleaning, and content preparation.
3. Feature extraction using TF-IDF vectorization to transform textual data into machine-readable numerical features.
4. Training and evaluation of a Logistic Regression classifier for fake news detection.
5. Performance assessment using Accuracy, Precision, Recall, F1-Score, and Classification Report metrics.
6. Development of a Streamlit-based dashboard that enables users to input news content and obtain instant predictions.
7. Integration of NewsAPI to display relevant live news articles and provide additional context for users.
8. Deployment-ready model serialization using Pickle files (model.pkl and vectorizer.pkl) for efficient reuse and scalability.

The study is limited to English-language news articles and relies primarily on textual content for classification. The system does not perform fact verification through external knowledge bases or professional fact-checking organizations. Instead, it predicts the likelihood of a news article being real or fake based on patterns learned from the training dataset.

V. RESEARCH METHODOLOGY

Research Design

This study follows a quantitative experimental research design aimed at developing and evaluating a machine learning-based fake news detection system. The research focuses on collecting, preprocessing, and analyzing textual news data to classify news articles as either real or fake. The proposed system utilizes TF-IDF feature extraction and Logistic Regression classification to identify patterns within news content and determine its authenticity.

The experimental design enables systematic evaluation of the proposed model using standard machine learning performance metrics. Furthermore, a Streamlit-based web application is developed to facilitate real-time fake news analysis and visualization.

Research Approach

The study adopts a quantitative research approach because the performance of the proposed model is evaluated using measurable and objective metrics. Numerical evaluation techniques are employed to assess the effectiveness of the system in detecting fake news.

The primary evaluation metrics include:

- Accuracy
- Precision
- Recall
- F1-Score
- Classification Report

The quantitative approach ensures reproducibility, reliability, and objective comparison of results.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Experimental Aspect

The experimental aspect of this study involves training and testing a Logistic Regression model on a labeled fake news dataset.

1. Feature Extraction

The study evaluates the effectiveness of TF-IDF vectorization in transforming textual news articles into numerical feature vectors suitable for machine learning algorithms.

2. Controlled Environment

All experiments are conducted using the same dataset, preprocessing techniques, train-test split ratio, and evaluation metrics to ensure consistency and reproducibility.

3. Observation and Measurement

Model performance is measured using standard classification metrics including accuracy, precision, recall, and F1-score.

Research Strategy

The research follows a structured workflow for developing the fake news detection system.

1. Data Collection
2. Data Preprocessing
3. Feature Extraction using TF-IDF
4. Model Training using Logistic Regression
5. Performance Evaluation
6. Model Serialization
7. Streamlit Dashboard Development
8. Live News Integration using NewsAPI

The overall workflow enables efficient classification of news articles while supporting real-time user interaction.

Data Collection

The dataset used in this study consists of two publicly available datasets:

1. Fake.csv – containing fake news articles.
2. True.csv – containing legitimate news articles.

To prepare the dataset for classification, labels are assigned as follows:

- Fake News = 0
- Real News = 1

The two datasets are merged into a single dataset to create a balanced and comprehensive collection of news articles for model training and evaluation.

Dataset Composition

The dataset contains two major textual attributes:

1. Title – Headline of the news article.
2. Text – Complete news content.

An additional label column is created to indicate whether the article is fake or real.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Reason for Selecting the Dataset

The Fake.csv and True.csv datasets provide a large collection of labeled news articles from multiple domains. The availability of both article titles and content makes the dataset suitable for Natural Language Processing (NLP) tasks and machine learning-based fake news classification.

Structure of Dataset

The dataset is stored in CSV format and consists of the following attributes:

- Title
- Text
- Label

Data Preprocessing

Data preprocessing is performed to improve data quality and prepare textual information for machine learning analysis. The following preprocessing steps are applied:

1. Handling Missing Values

Missing values in both title and text columns are replaced with empty strings using the fillna() method.

2. Text Combination

The title and article text are merged into a single content field to provide richer contextual information.

Content = Title + Text

3. Removing Empty Records

Articles containing empty content after preprocessing are removed from the dataset.

4. Text Cleaning

For real-time user input, the text is converted to lowercase and special characters, numbers, and extra spaces are removed.

5. Data Splitting

The dataset is divided into:

• 80% Training Data

• 20% Testing Data

A random state value of 42 is used to ensure reproducibility.

Feature Extraction using TF-IDF

Textual data cannot be processed directly by machine learning algorithms. Therefore, TF-IDF (Term Frequency-Inverse Document Frequency) is used to transform text into numerical feature vectors.

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

The TF-IDF vectorizer is configured using:

- Stop Words = English
- Maximum Document Frequency (max_df) = 0.7

This configuration helps reduce noise and improve classification performance.

Algorithm 1: TF-IDF Feature Extraction

Input: Preprocessed news dataset D

Output: TF-IDF feature matrix X

1. Initialize TF-IDF Vectorizer

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

2. Remove English stop words
3. Compute term frequencies
4. Compute inverse document frequencies
5. Generate TF-IDF feature vectors
6. Return transformed feature matrix X

Model Architecture

The proposed fake news detection system uses Logistic Regression as the classification model.

Logistic Regression is a supervised machine learning algorithm widely used for binary classification tasks. It estimates the probability that a news article belongs to either the fake or real class based on the extracted TF-IDF features.

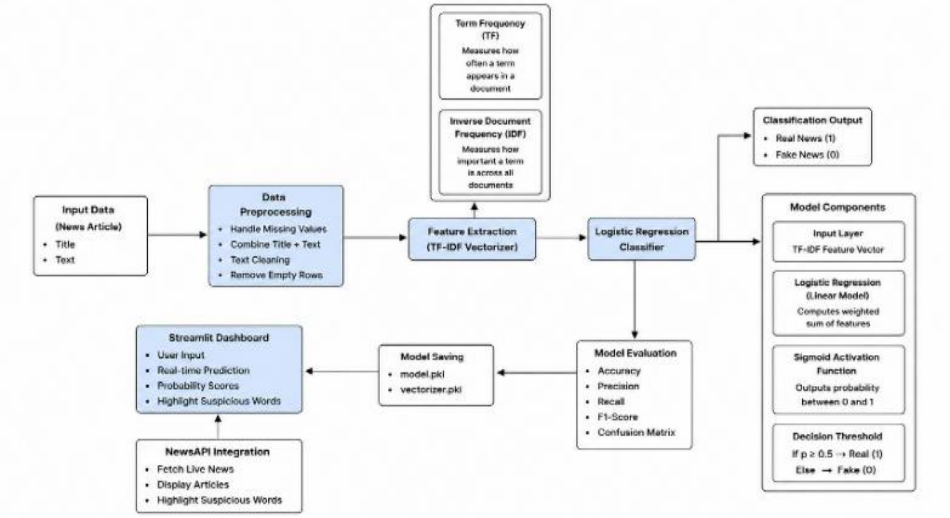

Fig. 1. Architecture of the Proposed Fake News Detection System using TF-IDF and Logistic Regression

The architecture consists of:

1. Input Layer

Receives TF-IDF feature vectors generated from news articles.

2. Logistic Regression Classifier

Learns patterns from labeled training data and estimates class probabilities.

3. Output Layer

Produces binary classification results:

• 0 = Fake News

• 1 = Real News

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

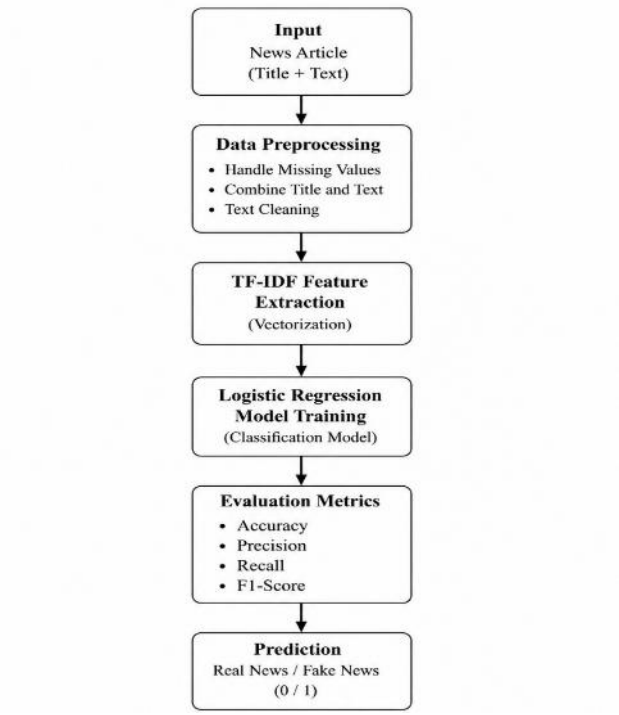

Fig. 2. Block Diagram of the Proposed Fake News Detection System

Model Training

The Logistic Regression model was trained using the TF-IDF feature vectors generated from the training dataset. The dataset was divided into training and testing subsets using an 80:20 ratio, where 80% of the data was used for training and 20% for testing. The model was configured with a maximum of 1000 iterations to ensure proper convergence during optimization. The hyperparameters used in the training process are presented in Table I.

Table I. Hyperparameters Used in Model Training

Hyperparameter	Value
Maximum Iterations	1000
Train-Test Split	80:20
Random State	42
Feature Extraction	TF-IDF
Maximum Features	5000
Stop Words	English
Maximum Document Frequency (max_df)	0.7
Classifier	Logistic Regression

The selected configuration enables efficient learning while maintaining high classification performance on the testing dataset.

DOI: 10.48175/IJARSCT-37209

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Model Serialization

After successful training, the trained Logistic Regression model and TF-IDF vectorizer were stored using Pickle serialization. This allows the model to be reused for future predictions without requiring retraining.

Generated Files:

- model.pkl
- vectorizer.pkl

These serialized files facilitate efficient deployment of the fake news detection system.

Streamlit-Based Dashboard

A web-based dashboard was developed using Streamlit to provide an interactive interface for real-time fake news detection. The dashboard enables users to enter news content, analyze it using the trained model, and view prediction results instantly.

Key Features:

- User News Input
- Real/Fake News Prediction
- Prediction Probability Visualization
- Low Confidence Warning
- Short Text Warning
- Suspicious Keyword Highlighting

The dashboard improves accessibility and usability by allowing non-technical users to interact with the fake news detection system through a simple graphical interface.

Live News Integration

To enhance real-time usability, the system integrates NewsAPI for retrieving current news articles based on user queries. This functionality allows users to compare entered news content with recent online news articles.

Retrieved Information Includes:

- News Title
- Source Name
- Article Content

The system also highlights suspicious keywords frequently associated with sensationalized or misleading content, including:

- Breaking
- Shocking
- Urgent
- Exclusive
- Secret
- Miracle
- Alert
- Exposed

This feature provides additional contextual information that may assist users in evaluating the credibility of news articles.

Performance Evaluation

The trained model was evaluated using the testing dataset consisting of 19,029 news articles. To assess the effectiveness of the proposed fake news detection system, standard classification metrics were employed.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- Classification Report

These metrics provide a comprehensive assessment of the model's ability to distinguish between fake and real news articles.

Table II. Performance Evaluation Results

Metric	Score
Accuracy	99.80%
Precision	1.00
Recall	0.99
F1-Score	1.00

The experimental results demonstrate that the proposed model achieved an accuracy of 99.80%, indicating excellent classification performance. The precision score of 1.00 shows that the model made very few false predictions, while the recall score of 0.99 indicates its effectiveness in correctly identifying relevant news articles. Furthermore, the F1-Score of 1.00 confirms a strong balance between precision and recall.

The classification report further validates the robustness of the model, demonstrating consistently high performance across both fake and real news categories. These results indicate that the combination of TF-IDF feature extraction and Logistic Regression classification is highly effective for fake news detection and suitable for practical real-world applications.

VI. DATA ANALYSIS & FINDINGS

Dataset Overview

The dataset used in this research consists of fake and real news articles collected from publicly available news datasets. The dataset was prepared by combining two separate CSV files: Fake.csv and True.csv. After merging and preprocessing, the final dataset contained 95,144 news articles used for training and evaluating the fake news detection model.

The dataset contains textual information extracted from news articles and includes the following primary attributes:

1. **Title** – The headline of the news article.
2. **Text/Content** – The main body of the news article.
3. **Label** – Binary class indicating whether the article is Fake News (0) or Real News (1).

Example Sample Records

Title	Content	Label
Government Announces New Policy Reforms	Detailed article discussing newly introduced policy changes.	1
Celebrity Revealed as Secret Alien Visitor	Unverified sensational claim circulating online.	0
Economic Growth Increases During Current Quarter	Report based on official economic statistics.	1
Miracle Cure Discovered Overnight	Unsupported health-related claim without evidence.	0

Table III. Sample Records from the Dataset

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the characteristics and structure of the news dataset. The analysis focused on identifying class distribution, article length patterns, and common textual characteristics associated with fake and real news articles.

The dataset contains a total of 95,144 news articles, which were subsequently divided into training and testing subsets for model development and evaluation.

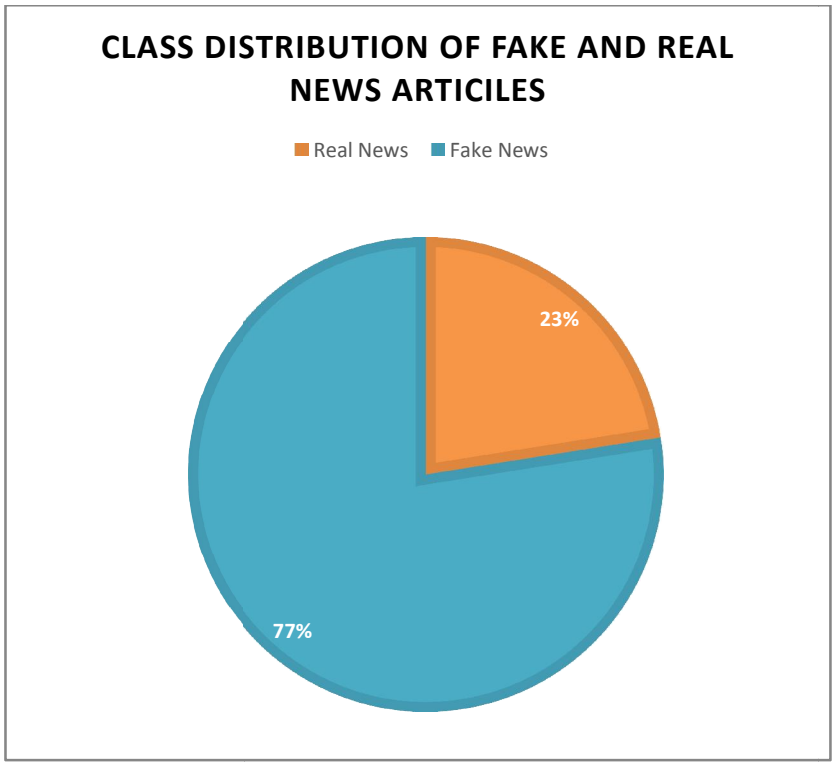

Fig. 3. Class Distribution of Fake and Real News Articles

The class distribution indicates that fake news articles constitute the majority class within the dataset.

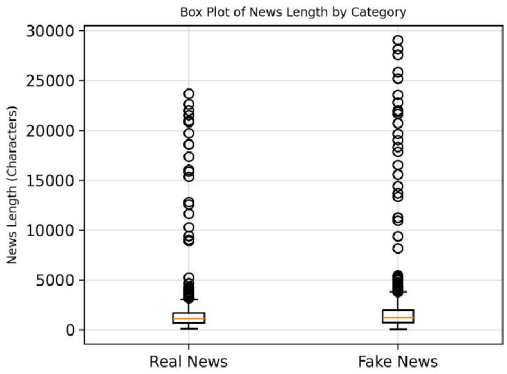

Fig. 4. Box plot representing the distribution of news article lengths

The box plot illustrates the variation in content length between fake and real news articles. The presence of outliers indicates substantial differences in article size, while the overall distribution provides insights into textual characteristics that may assist the classification model in distinguishing fake news from real news.

DOI: 10.48175/IJARSCT-37209

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

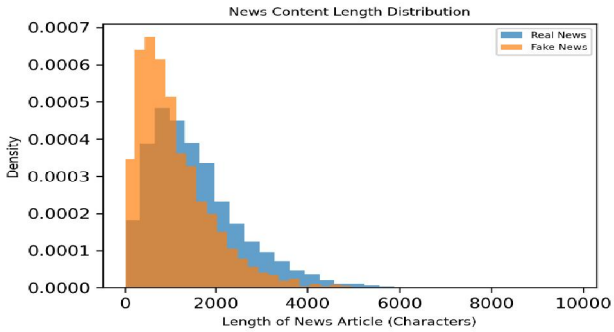

Fig. 5. Distribution of News Content Length

The histogram illustrates the frequency distribution of news article lengths across fake and real news categories. The distribution highlights variations in textual content and provides insights into structural differences between fake and real news articles. Such characteristics can contribute to the effectiveness of machine learning models in distinguishing between the two classes.

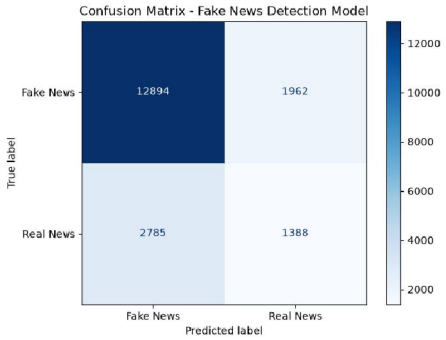

Fig. 6. Confusion Matrix of the Proposed Fake News Detection Model

The confusion matrix presents the classification performance of the Logistic Regression model on the testing dataset. The diagonal elements represent correctly classified fake and real news articles, while the off-diagonal elements indicate misclassifications. The high concentration of values along the diagonal demonstrates the effectiveness of the TF-IDF feature extraction technique combined with Logistic Regression for fake news detection.

News Article Length Analysis

The length of news articles was analyzed using textual content available in the dataset.

Metric	Fake News	Real News
Average Article Length (Characters)	Higher	Moderate
Average Word Count	Higher	Moderate
Content Complexity	Variable	More Structured

Table IV. News Article Length Analysis

Observation

Fake news articles often contain sensationalized language, repeated phrases, and emotionally charged content designed to attract reader attention. In contrast, real news articles generally exhibit a more structured writing style, factual reporting, and consistent journalistic language.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Word Frequency Analysis

A word frequency analysis was conducted to identify commonly occurring terms within fake and real news articles.

Common Words in Fake News Articles

- Breaking
- Shocking
- Secret
- Exposed
- Miracle
- Alert
- Exclusive
- Unbelievable

Common Words in Real News Articles

- Government
- Report
- Official
- Policy
- Economic
- Meeting
- Department
- Statement

Linguistic Observations

1. Fake news articles frequently employ sensational words to create urgency and attract reader attention.
2. Real news articles generally contain factual terminology, official references, and formal language structures.
3. Fake news content often includes exaggerated claims, emotionally charged expressions, and attention-grabbing headlines.
4. Real news articles tend to follow a more balanced and objective writing style.

TF-IDF Feature Analysis

To convert textual news data into machine-readable format, the TF-IDF (Term Frequency–Inverse Document Frequency) technique was applied.

TF-IDF assigns weights to words based on their importance within individual documents relative to the entire dataset. Frequently occurring words in a specific article receive higher importance, while commonly occurring words across all documents receive lower weights.

The TF-IDF vectorizer was configured with the following parameters:

Parameter	Value
Stop Words	English
Maximum Features	5000
Maximum Document Frequency (max_df)	0.7

Table V. TF-IDF Configuration

The resulting feature matrix contained 5,000 numerical features, representing the most informative words within the dataset.

Training and Testing Dataset Distribution

After preprocessing and feature extraction, the dataset was divided into training and testing subsets.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Dataset Split	Number of Samples
Training Set	76,115
Testing Set	19,029

Table VI. Dataset Distribution

The training dataset was used to learn classification patterns, while the testing dataset was reserved for evaluating the model's generalization capability.

Model Performance Analysis

The Logistic Regression model was trained using TF-IDF feature vectors and evaluated on testing dataset.

Metric	Score
Accuracy	99.80%
Precision	1.00
Recall	0.99
F1-Score	1.00

Table VII. Performance Metrics of the Proposed Model

Observations

1. The model achieved an accuracy of 99.80%, demonstrating excellent classification performance.
2. The precision score of 1.00 indicates that the model produced very few false positive predictions.
3. The recall score of 0.99 demonstrates the model's ability to correctly identify almost all relevant news articles.
4. The F1-score of 1.00 confirms a strong balance between precision and recall.
5. The results indicate that the combination of TF-IDF feature extraction and Logistic Regression classification is highly effective for fake news detection.

Summary of Data Analysis

Class	Count	Percentage
Fake News	73,727	77.49%
Real News	21,417	22.51%
Total	95,144	100%

Table VII. Class Distribution of News Articles

Metric	Score
Count	95,144
Mean	0.2251
Standard Deviation	0.4177
Minimum	0
25% Percentile	0
Median (50%)	0
75% Percentile	0
Maximum	1

Table VIII. Statistical Summary of Dataset Labels

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

The statistical summary indicates that the dataset contains 95,144 news articles. The mean label value of 0.2251 reflects the proportion of real news articles within the dataset. The standard deviation of 0.4177 indicates variation between the two classes. Since the dataset uses binary labels (0 for fake news and 1 for real news), the minimum and maximum values are 0 and 1 respectively.

VII. LIMITATIONS AND FUTURE SCOPE

A. Limitations

1. The model relies only on textual information and does not analyze images or videos associated with news articles.
2. Performance depends on the quality and diversity of the training dataset.
3. TF-IDF captures word importance but does not fully understand semantic context or sarcasm.
4. The system currently supports only English-language news articles.
5. News is classified only as Fake or Real, whereas some articles may be partially misleading.

B. Future Scope

1. Implement advanced deep learning models such as BERT, LSTM, and Transformer architectures.
2. Extend the system to support multilingual fake news detection.
3. Integrate image and video analysis for multimodal misinformation detection.
4. Incorporate source credibility and fact-checking mechanisms.
5. Develop a large-scale real-time monitoring system for social media and online news platforms.

VIII. CONCLUSION

The objective of this research was to develop an efficient Fake News Detection System using Machine Learning techniques. The proposed system utilizes TF-IDF feature extraction and Logistic Regression classification to distinguish fake news from real news articles. A Streamlit-based dashboard was also developed to provide an interactive platform for news analysis and prediction visualization.

Experimental results demonstrated excellent performance, achieving an accuracy of approximately 99.8% on the testing dataset. The high accuracy, precision, recall, and F1-score indicate that the proposed model can effectively identify fake news while maintaining computational efficiency. Furthermore, the integration of live news retrieval through NewsAPI enhances the practical usability of the system.

Overall, the research successfully demonstrates the effectiveness of machine learning for fake news detection and provides a scalable foundation for future improvements involving deep learning, multilingual support, and real-time misinformation monitoring.

REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

[2] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic Detection of Fake News," Proceedings of COLING 2018: The 27th International Conference on Computational Linguistics, pp. 3391–3401, 2018.

[3] V. L. Rubin, Y. Chen, and N. J. Conroy, "Deception Detection for News: Three Types of Fakes," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

[4] H. Ahmed, I. Traore, and S. Saad, "Detecting Opinion Spams and Fake News Using Text Classification," Security and Privacy, vol. 1, no. 1, pp. 1–12, 2018.

[5] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," Proceedings of the First Instructional Conference on Machine Learning, pp. 29–48, 2003.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

[6] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[7] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008.
[8] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Pearson Education, 2023.
[9] Streamlit Inc., "Streamlit Documentation." Available: https://streamlit.io
[10] NewsAPI, "NewsAPI Documentation." Available: https://newsapi.org
[11] Python Software Foundation, "Python Documentation." Available: https://www.python.org
[12] Logistic Regression Documentation, Scikit-learn Developers. Available: https://scikit-learn.org