A Project Synopsis On

FAKE NEWS DETECTION

Submitted by:

1] Dhanashri Ekanath Khondre [Roll No. 29]

2] Sakshi Sudhir Mohite [Roll No. 42]

3] Aditi Rajiv Patil [Roll No. 47]

4] Bhumika Sandip Patil [Roll No. 48]

Under The Guidance Of

Prof. Supriya Laykar

(Associate Professor**)**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING [DATA SCIENCE]

Dr. D. Y. Patil Pratishthan's College Of Engineering
Salokhenagar, Kolhapur

YEAR 2024-25

INDEX

- **Introduction :**

Fake news detection is an essential problem in today's digital world, where misinformation spreads quickly. This project aims to develop a **Fake News Prediction System** using **Machine Learning**, which classifies news articles as either real or fake based on their textual content.

The internet has become a major source of news consumption. However, the ease of sharing content online has led to an increase in the spread of fake news—false or misleading information presented as news. Fake news can influence public opinion, cause panic, and undermine trust in legitimate news outlets.

To address this, our project builds a **machine learning-based system** that can automatically detect whether a piece of news is real or fake by analysing the text content. The system uses **Natural Language Processing (NLP)** techniques and **classification algorithms** to analyze news articles and make predictions. This project aims to help platforms, journalists, and users quickly identify misinformation and take action against it.

- **Literature Survey / Review of Existing System (Detailed) :**

Overview :

The problem of fake news detection has gained significant attention over the past few years. Several systems and models have been proposed, each attempting to solve the problem using a different approach.

2.1 Existing Approaches :

Rule-Based Systems

These systems rely on predefined rules and patterns to classify text as fake or genuine. Although easy to implement, they lack the flexibility and adaptability to cope with the evolving nature of fake content.

Traditional Machine Learning Models

Support Vector Machines (SVM), Naive Bayes, and Logistic Regression are commonly used for fake news detection. These models rely on manually engineered features extracted using NLP techniques.

Deep Learning Models

Advanced architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM), have been employed to improve accuracy by learning hierarchical representations of text.

Hybrid approaches

Combining traditional ML models with rule-based systems and deep learning has shown improved results in terms of accuracy and robustness. These models aim to utilize the strengths of different techniques.

2.2 Limitations of Existing Systems

- Most systems are limited to English-language datasets.
- Difficulty in handling sarcasm, satire, and ambiguous text.
- High dependency on labeled datasets.
- Inability to adapt to new forms of fake content, such as AI-generated news.

- **Problem Specification / Proposed System**

3.1 Problem Statement

To design and implement an intelligent fake news detection system that automatically classifies news content as fake or real using machine learning algorithms and natural language processing.

3.2 Objectives

- Automate the process of identifying fake news articles.
- Improve classification accuracy through pre-processing and vectorization.
- Use logistic regression for real-time binary classification.
- Evaluate the model with standard metrics.

3.3 Proposed Methodology

Data Collection

The dataset used is a labelled CSV file containing real and fake news headlines and bodies.

Data Pre-processing

- Removing punctuations and special characters.
- Converting all text to lowercase.
- Removing stopwords (e.g., "and", "the", "is").
- Applying stemming or lemmatization.

Feature Extraction

Term Frequency-Inverse Document Frequency (TF-IDF) is used to convert text into numerical feature vectors.

Model Training

- Logistic Regression classifier is used to learn from the extracted features.
- The model is trained on a subset of data and evaluated using the remaining data.

Evaluation

Confusion matrix, accuracy, precision, recall, and F1-score are used for model evaluation.

- **Hardware / Software Requirements :**

4.1 Hardware Requirements

- Processor: Intel i5 or higher
- RAM: 8GB minimum (16GB recommended)
- Storage: At least 50 GB free
- GPU: Optional for deep learning tasks

4.2 Software Requirements

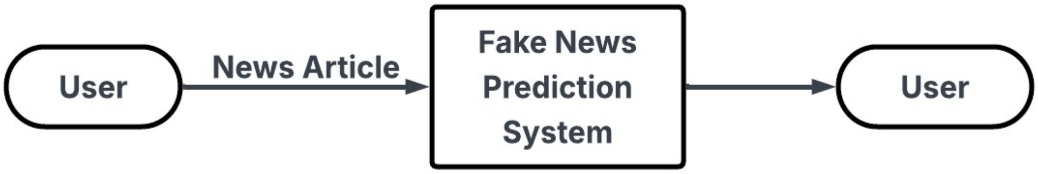
- Operating System: Windows 10/11, Linux (Ubuntu recommended)
- Programming Language: Python 3.8+
- IDE: Jupyter Notebook, VS Code
- Libraries:
  - NLP: NLTK, SpaCy
  - ML: Scikit-learn
  - Data Handling: Pandas, NumPy
- Version Control: Git, GitHub

4.3 Development Environment

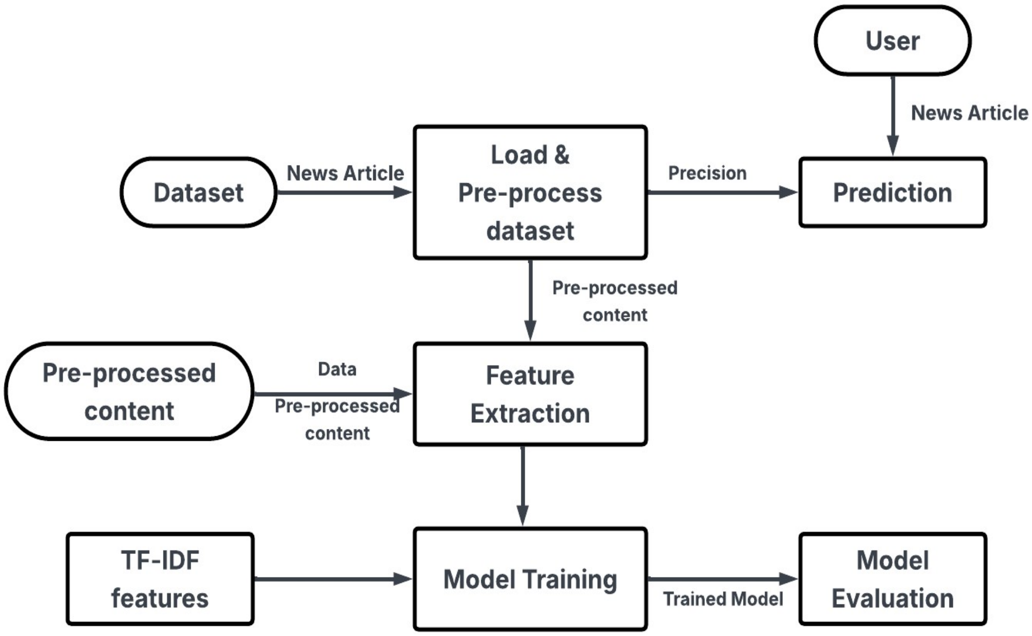
- Anaconda (for environment management)
- Jupyter Notebook for implementation
- Colab

- **System Design and Implementation**

DFD Level 0 :



DFD Level 1 :



- **Algorithm :**

1. **User uploads a dataset** of news articles (real + fake).
2. **Text is cleaned:** remove punctuation, lowercase conversion, stopwords removal, and stemming.
3. **Data split** into training and testing (usually 80:20).
4. **TF-IDF Vectorizer** transforms text into numerical format for machine learning.
5. **Logistic Regression model** is trained on the data.
6. **Model predicts** whether test articles are real or fake.
7. **Accuracy score and confusion matrix** evaluate the model's performance.

Data Pre-processing:

Lowercasing, stopword removal, tokenization, lemmatization

Feature Extraction

TF-IDF vectorizer converts text into numerical values

Model Training

Logistic Regression trained using 80/20 train-test split

Prediction and Evaluation

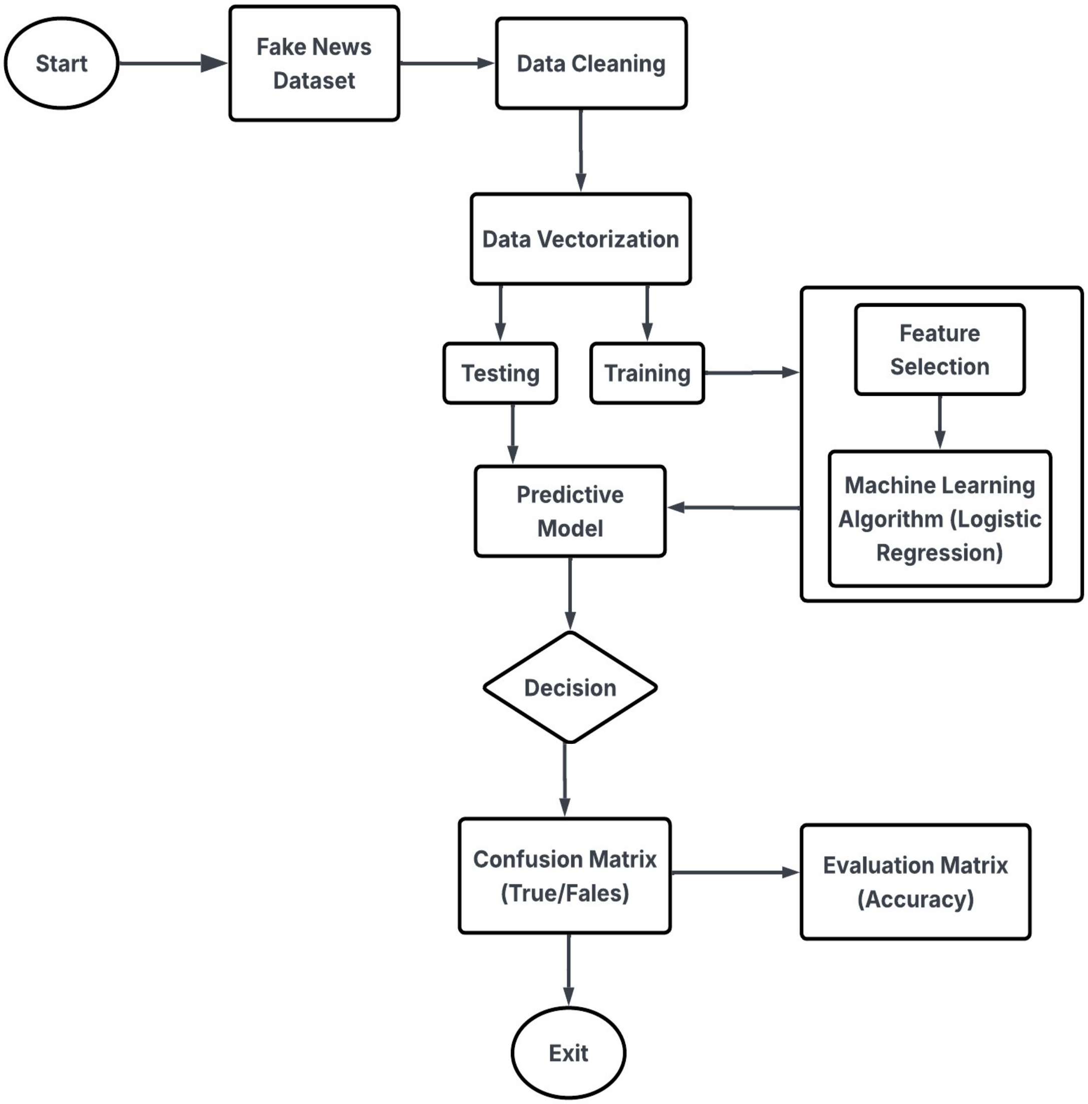
- Model used to predict new samples
- Evaluation with accuracy and confusion matrix

Model Results (Based on Code)

- Accuracy: 98.3%
- Precision: 98%

- Recall: 98%
- F1 Score: 98%

- **Flowchart:**



- **Conclusion:**

The fake news detection system effectively classifies news as real or fake using text classification techniques. The integration of NLP pre-processing and TF-IDF with a logistic regression classifier provides a reliable and interpretable solution.

Through evaluation metrics, the model demonstrates high accuracy, making it suitable for real-world deployment. Further improvements can be achieved by using larger datasets, multilingual support, or integrating deep learning models.

- **Reference / Biography**

1. Zhou, Xinyi, and Reza Zafarani. "Fake News: A Survey of Research, Detection Methods, and Opportunities." (2018)
2. Shu, Kai, et al. "Fake News Detection on Social Media: A Data Mining Perspective." (2017)
3. Scikit-learn documentation: https://scikit-learn.org
4. Kaggle datasets: https://www.kaggle.com
5. NLTK documentation: https://www.nltk.org