Task 6: News Article Classification

1. Introduction: Fake news involves the deliberate creation and spread of false or misleading articles, often disguised as legitimate news, to deceive readers, influence opinions, or generate profit. It spreads rapidly online, eroding public trust in media, distorting facts on critical issues like politics or health, and causing real-world harm. Combating it requires critical thinking, verifying sources, and supporting credible journalism.

2. Abstract: This project develops an automated system to detect fake news articles using AI and machine learning techniques. By analyzing linguistic patterns, source credibility, and content consistency, the tool aims to identify potentially false or misleading information rapidly. It provides users with real-time verification support, enhancing media literacy and aiding in the critical evaluation of online news sources to combat the spread of digital misinformation. The solution prioritizes efficiency, accuracy, and user-friendly integration into existing platforms.

3.Tools used:

- **Core Libraries**:
    - pandas & numpy: For data loading, manipulation, and preprocessing
    - scikit-learn: For machine learning operations including:
        - TfidfVectorizer: Text vectorization (TF-IDF)
        - LogisticRegression: Classification model
        - train_test_split: Data splitting
        - Evaluation metrics (classification_report)
- **Text Processing**:
    - re (Regular Expressions): For text cleaning and pattern matching
- **Model Persistence**:
    - joblib: For saving/loading the trained model and vectorizer
- **Deployment**:
    - streamlit: For building the web interface and interactive demo

4. Steps Involved:

i. Data Preparation:

- Loaded datasets: True.csv (real news) and Fake.csv (fake news)
- Added labels: 1 for real news, 0 for fake news
- Combined datasets and shuffled rows randomly

- Removed irrelevant columns (title, subject, date)

ii. Text Preprocessing

- Cleaned text using wordopt() function:
    - Lowercasing
    - Removed URLs, HTML tags, punctuation, digits, and newlines
- Applied cleaning to all text in the text column

iii. Feature Engineering

- Split data into features (X: processed text) and labels (y)
- Used TF-IDF Vectorization to convert text to numerical features
- Created train/test splits (70% training, 30% testing)

iv. Model Training

- Initialized Logistic Regression classifier
- Trained model on vectorized training data (xv_train, y_train)

v. Evaluation

- Made predictions on test set (xv_test)
- Calculated accuracy score
- Generated classification report (precision, recall, f1-score)

vi. Deployment

- Saved model and vectorizer using joblib
- Built web interface with Streamlit:
    - Text input for news content
    - Real-time prediction display
    - Visual feedback (success/error messages)

5. Conclusion: This project developed a streamlined fake news detector using logistic regression on TF-IDF vectorized text data, achieving high accuracy in classifying articles. Key steps included: preprocessing news text (cleaning URLs/punctuation), training the model on labeled datasets, evaluating performance metrics, and deploying an interactive Streamlit web app for real-time verification. The solution demonstrates an effective ML pipeline for combating misinformation, with future potential for integrating deep learning or multilingual support.