FAKE NEWS DETECTION SYSTEM

Fake News Detection is a natural language processing task that involves identifying and classifying news articles or other types of text as real or fake. The goal of fake news detection is to develop algorithms that can automatically identify and flag fake news articles, which can be used to combat misinformation and promote the dissemination of accurate information.

DATASET

There are two datasets, one containing real data and the other containing fake data. Both datasets have four columns: 'title', 'text', 'subject', and 'date'.

Title: Contains the title of the news articles.
Type: Object

Text: Contains the main text or content of the news articles.
Type: Object

Subject: Indicates the subject or category of the news articles.
Type: Object

Date: Represents the date when the news articles were published.
Type: Object

Steps followed

- Importing Libraries and Datasets
- Data Cleaning
- EDA
- Data Preprocessing
- Preprocessing and analysis of News column
- Converting text into Vectors
- Model training, Evaluation, and Prediction

DATA CLEANING

Data cleaning process was applied to ensure the integrity and quality of the datasets containing real and fake news articles. The following steps were performed:

Handling Null Values:

- The dataset was examined for the presence of null values in each column ('title', 'text', 'subject', 'date').
- Null values, if any, were identified and addressed appropriately. In this case, it was observed that there were no null values in any of the columns, ensuring completeness and consistency in the data.

Duplicate Removal:

- Duplicates within the datasets were examined to eliminate redundancy and maintain data accuracy.
- Any duplicate entries, if present, were identified and removed, ensuring that each record in the dataset is unique. This step contributes to preventing bias and inaccuracies in subsequent analyses or model training.

EDA

- There are 23,478 entries in the fake news dataset, indicating the total number of news articles.
- The date May 10, 2017, is the most common date, with 46 occurrences.
- The subject "News" is the most common, appearing 9,050 times in the dataset.
- There are 21,211 entries in this real news dataset.
- The subject "politicsNews" is the most common, appearing 11,220 times.
- The date December 6, 2017, is the most common, with 166 occurrences.
- Longer titles in fake news might be used to grab attention and provoke emotional responses.
- Shorter text lengths in fake news could suggest a lack of substance or credible information's
- The subject "News" is the most common, appearing 9,050 times in the dataset.
- The subject "politicsNews" is the most common in fake news scenario, appearing 11,220 times.
- The graph suggests that politics is a common subject for fake news articles

DATA PREPROCESSING

The data preprocessing phase involved several crucial steps to prepare the datasets for subsequent analysis and modelling. The key preprocessing steps include:

Data Splitting:

The datasets were split into training, testing, and validation sets to facilitate model training, evaluation, and tuning.

Column Operations:

- Certain columns were merged to consolidate information and improve feature representation.
- Unwanted columns that did not contribute to the analysis were removed for simplicity and efficiency.

Labelling Data:

- A new column named 'class' was introduced to distinguish between real and fake news articles. This categorical label serves as the target variable for subsequent classification tasks.

Text Cleaning Using NLTK Libraries:

- NLTK libraries were employed to clean and preprocess the text data.
- Operations included converting text to lowercase, removing links, brackets, and other unwanted characters to enhance the quality of textual information.

Text Tokenization and Stopword Removal:

- The cleaned text was tokenized into individual words to facilitate further analysis.
- Common English stopwords were removed to focus on significant content and reduce noise.

Vectorization:

- Two vectorization techniques, CountVectorizer and TF-IDF Vectorizer, were applied to convert the text data into numerical format.
- CountVectorizer transformed text into a bag-of-words representation, while TF-IDF Vectorizer considered term frequency and inverse document frequency to represent the importance of words.

MODELS USED FOR TRAINING

Logistic Regression:

- **Suitability**: Chosen for its simplicity and efficiency in binary classification tasks.
- **Characteristics**: Logistic regression models the probability of a binary outcome using a linear combination of input features, providing interpretable results.

Naive Bayes:

- **Suitability**: Selected due to its effectiveness in text classification tasks and simplicity.
- **Characteristics**: Naive Bayes is based on Bayes' theorem, assuming independence between features, making it efficient for text data.

Random Forest:

- **Suitability**: Chosen for its ensemble learning capability and robustness in handling complex relationships in data.
- **Characteristics**: Random Forest combines multiple decision trees to improve predictive accuracy and control overfitting.

Performance Metrics: The models were evaluated using various metrics, including accuracy, precision, recall, and F1-score, to assess their effectiveness in making predictions on both real and fake news.

Logistic Regression:

- The logistic regression model achieved high precision (99%) for non-fake news and high recall (98%) for fake news.
- It demonstrated balanced F1-scores (0.99 for non-fake, 0.98 for fake), showcasing a harmonious trade-off between precision and recall.
- The overall accuracy of 98% indicates effective classification of both fake and non-fake news instances.

Naive Bayes:

- The model achieves a balanced performance with similar precision, recall, and F1-score for both classes (0 and 1).
- The overall accuracy of 93% indicates a reliable classification across both fake and non-fake news instances.
- Slightly higher precision (94%) for predicting fake news suggests accurate identification of actual instances.
- Both precision (93%) and recall (94%) are high, indicating effective classification of non-fake news instances.

Random Forest:

- An accuracy of 99% indicates highly reliable classification across the dataset.
- Class 0 (non-fake news) and Class 1 (fake news) both show exceptional precision (99%) and recall (99%), emphasizing the model's proficiency in distinguishing between the two.

MODEL TESTING

The manual testing function is designed to evaluate the authenticity of a given news article using three different machine learning models: Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF). The function processes the input news text, applies the necessary preprocessing steps, and predicts whether the article is fake or real based on the patterns learned during the training phase.

MODEL EVALUATION

- All models demonstrate effectiveness in fake news detection, with each having its strengths.
- Logistic Regression and Random Forest outperform Naive Bayes in terms of precision, recall, and overall accuracy.
- Random Forest stands out with exceptional performance across all metrics, providing a robust solution for fake news classification.

CONCLUSION

The implemented machine learning for fake news detection involves robust data preprocessing, training, and testing. Utilizing Logistic Regression, Naive Bayes, and Random Forest models, the system demonstrates high accuracy and reliability in distinguishing between real and fake news. The manual testing function allows for interactive assessment, providing practical insights into the models' predictions for real-world use cases. This comprehensive approach offers an effective solution for combating misinformation and contributes valuable tools for news authenticity verification.