Project Summary and Approach

In this project, we aimed to build a **news classification model** to distinguish between **True** and **Fake** news articles. The primary goal was to design a system that could automatically identify the authenticity of news based on its textual content. The approach involved **text processing**, **feature extraction**, and the application of various **machine learning algorithms**.

Text Processing and Feature Engineering

We began by cleaning the dataset, which involved removing unnecessary elements like special characters, numbers, and stopwords, as well as ensuring that all text was in lowercase for uniformity. Key techniques used in text processing included:

1. **Tokenization**: Breaking down the text into individual words to understand the structure of the content.
2. **Lemmatization**: Reducing words to their base form (e.g., "running" to "run") to standardize the text and reduce redundancy.
3. **Part-of-Speech Tagging**: Identifying the parts of speech for each word in the text, which helped in filtering for the most meaningful words (nouns) related to the topic.
4. **Word Embeddings**: We utilized **Word2Vec** to convert words into numerical vectors that capture their semantic meaning. This approach helped the model understand the relationships between different words and their contexts.

We also performed **n-gram extraction** (unigrams, bigrams, and trigrams) to capture common word pairs and triplets, which can reveal important patterns in both true and fake news articles.

Data Preparation

After processing the text, we created a new DataFrame that combined the cleaned and lemmatized text. This was used for training and evaluation. The dataset was split into **70% training data** and **30% validation data** using **train_test_split**. This ensured that we had a robust model that could generalize well to unseen data.

Machine Learning Models

We experimented with several machine learning algorithms to build the classification model:

1. **Logistic Regression**: A simple and effective algorithm that we used as a baseline for binary classification.
2. **Decision Tree**: A non-linear model that splits the data based on features, offering interpretability but sometimes prone to overfitting.
3. **Random Forest**: An ensemble method that combines multiple decision trees to improve accuracy and generalization.

Each of these models was trained on the training data and evaluated using the **validation set**. We calculated performance metrics such as **accuracy**, **precision**, **recall**, and **F1-score** to assess the models.

Key Findings

Through the analysis, we identified some key patterns in the data:

1. **True News**: The articles classified as true news generally focused on factual information such as places, dates, and well-known events. The language used was neutral and informative.

2. **Fake News**: Fake news often contained more emotional, sensational language. It also frequently included polarizing terms that might appeal to readers' emotions rather than facts.

Model Performance

Out of all the models tested, the **Random Forest model** achieved the highest **F1-score**, indicating its ability to balance precision and recall better than others. Random Forest's performance was superior to that of **Logistic Regression** and **Decision Tree** models due to its ability to capture complex patterns in the data and prevent overfitting.

The **Random Forest** model, with its ensemble approach, was able to generalize better to unseen data, resulting in improved prediction accuracy and reduced variance compared to the simpler models.

Conclusion

By leveraging **semantic classification techniques** such as **text processing**, **lemmatization**, and **Word2Vec** embeddings, we were able to effectively analyze the textual content of news articles and differentiate between true and fake news. The **Random Forest model** demonstrated the best performance, providing a reliable way to classify news articles based on their content.

This approach can be beneficial in combating the spread of **fake news** by automating the classification process. It offers a scalable solution that could be applied to large volumes of news data to flag potentially fake articles for further review.