PROJECT REPORT

Title : Fake News Detection

Done By : Kishore Kumar B N

INTRODUCTION :

Fake news has become a prevalent issue in today's information-driven world. With the increasing spread of false information through social media and news outlets, it is essential to develop tools to detect fake news automatically. This project uses a machine learning approach to classify news articles as either Real or Fake using appropriate machine learning algorithms.

OBJECTIVE :

The goal of this project is to build a model that can accurately classify news articles into Real or Fake categories based on the text content of the article.

DATASET :

The dataset used for this project is the "Fake or Real News" dataset ,whixh indeed is a real life dataset souced from www.kaggle.com .

The dataset contained 6335 rows and 3 columns which were :

- Title : The title of the news article.
- Text : The body of the news article.
- Label : Each Of The Rows Were Labeled Whether The News Is *Real* Or *Fake*.

DATA PREPROCESSING :

1. Analysing the data : The original dataset was analysed to make the consideration of the columns that would be necessary for this project.
2. Neglecting Unnecessary Columns: The dataset contains an unnecessary 'Unnamed: 0' column which was not considered
3. Handling Missing Data : No missing data was found in the dataset.
4. Splitting the Data : The dataset was split into training and testing sets using a 80-20 split. This helps in evaluating the model's performance on unseen data.
5. Feature Extraction : The TF-IDF Vectorizer was used to convert the text of the news articles into numerical features, making it possible for the machine learning model to process and learn from the data. Stop words were removed to improve efficiency, and the 'max_df' parameter was set to 0.7 to ignore terms that appear in more than 70% of the articles.

MODEL SELECTION :

For this project, the **Multinomial Naive Bayes** algorithm was chosen. This algorithm is well-suited for text classification problems due to its simplicity and effectiveness with high-dimensional data like text.

MODEL TRAINING :

The model was trained using the training set, where the text column was used as the input, and the label column was used as the target variable.

1. The 'TF-IDF Vectorizer' was applied to transform the text data into numerical vectors.
2. The 'Naive Bayes Classifier' was then trained on the TF-IDF-transformed training data.

MODEL EVALUATION :

After training, the model was evaluated using the test set. The following metrics were used to assess the model's performance:

1. Accuracy : The percentage of correctly classified news articles.
2. Precision, Recall, F1-Score : These metrics were used to evaluate the balance between true positives and false positives, especially important for imbalanced data.

RESULTS :

- **Accuracy**: 84.53%
- **Classification Report**:

Label	Precision	Recall	F1-Score	Support
FAKE	0.97	0.73	0.83	923
REAL	0.79	0.98	0.88	978

- **Overall Accuracy**: 86%

The model performed well, with an F1-score of 0.83 for detecting fake news and 0.88 for detecting real news. The precision and recall values indicate that the model is better at detecting real news but still performs reasonably well in identifying fake news.

USER INPUT CLASSIFICATION :

To allow users to input their own news articles and classify them as *REAL* or *FAKE*, an interface was added where the user can enter a piece of news text, and the model will predict the label using the trained Naive Bayes model.

CONCLUSION :

This project successfully implemented a Fake News Detection system using a Multinomial Naive Bayes classifier. The model achieved an accuracy of 86% on the test data, showing that it can effectively distinguish between real and fake news articles.