PROJECT DOCUMENTATION

Introduction

This code helps you detect fake news by comparing two CSV files, one containing true news articles and the other containing fake news articles. It works by first reading the two files into dataframes. Then, it identifies the subjects that are only present in one dataset and plots the results in a bar chart. This can help you identify subjects that are commonly used in fake news articles.

About Dataset

The dataset for the code is a collection of CSV files, where each CSV file contains news articles. The CSV files should have the same column structure, with the first column containing the subject of the article.

The dataset can be divided into two subsets: the true news dataset and the fake news dataset. The true news dataset contains news articles that have been verified to be factual. The fake news dataset contains news articles that have been verified to be false or misleading.

The dataset can be used to train a machine learning model to detect fake news. The model can be trained to identify the features of fake news articles, such as the use of certain keywords or phrases, or the presence of grammatical errors. Once the model is trained, it can be used to predict whether a new news article is likely to be true or fake.

The dataset can also be used to study the nature of fake news. For example, researchers can use the dataset to identify the most common subjects of fake news articles, or to track how the spread of fake news changes over time.

Preprocessing steps

- **Remove punctuation and special characters: This can be done using regular expressions.**
- **Convert the text to lowercase: This helps to normalize the text and makes it easier for the machine learning model to process.**

- Tokenize the text: This involves splitting the text into individual words or tokens.
- Remove stop words: Stop words are common words that do not add much meaning to the text, such as "the", "is", and "of". Removing stop words can help to improve the performance of the machine learning model.
- Lemmatize or stem the words: Lemmatization and stemming are two techniques that can be used to reduce words to their root form. This can help to improve the performance of the machine learning model, especially if the dataset is small.

Choice of ML Algorithm:

- Logistic regression: Logistic regression is a type of supervised learning algorithm that can be used for classification tasks. Logistic regression works by fitting a logistic function to the data points. Logistic regression is a good choice for fake news detection because it is simple to implement and interpret.

Model Training:

1. Prepare the dataset: This involves preprocessing the dataset, such as removing punctuation and special characters, converting the text to lowercase, and tokenizing the text.
2. Split the dataset into training and testing sets: This is done to evaluate the performance of the model on unseen data. A common split is to use 80% of the data for training and 20% of the data for testing.
3. Choose a machine learning algorithm: There are a variety of machine learning algorithms that can be used for fake news detection, such as support vector machines, logistic regression, random forests, and neural networks.
4. Train the model: This involves feeding the training data to the machine learning algorithm and allowing it to learn the relationships between the features and the target variable.
5. Evaluate the model: This involves feeding the testing data to the trained model and measuring its accuracy.

Evaluation Metrics:

- Accuracy: This is the percentage of correctly predicted examples.
- Precision: This is the percentage of positive predictions that are correct.
- Recall: This is the percentage of actual positive examples that are correctly predicted.
- F1 score: This is a harmonic mean of precision and recall.

Result:

The code can be used to detect fake news by comparing two CSV files, one containing true news articles and the other containing fake news articles. The script works by first reading the two datasets into DataFrames. Then, it identifies the subjects that are only present in one dataset and plots the results in a bar chart. This can help to identify subjects that are commonly used in fake news articles.