Project: Fake News Detection Using NLP - Phase 3 Guidelines

Development Part 1: Loading and Preprocessing the Dataset

Dataset Collection:

Obtain a well-labeled fake news dataset from reputable sources or datasets like the "Fake News Challenge" dataset.

Data Loading:

Utilize a programming language like Python and libraries (e.g., Pandas) to load the dataset into your project.

Data Exploration:

Conduct exploratory data analysis to understand the dataset's characteristics.

Check for missing data, class distribution, and other relevant statistics.

Text Preprocessing:

Perform the following preprocessing steps:

Convert all text to lowercase.

Remove punctuation, special characters, and numerical values.

Tokenization:

Split text into individual words or tokens.

Stopword Removal :

Eliminate common words.

Lemmatization or stemming to reduce words to their base form.

Text Vectorization:

Convert preprocessed text data into numerical form using techniques like TF-IDF or word embeddings (Word2Vec, GloVe).

Train-Test Split:

Split the dataset into training and testing sets to facilitate model training and evaluation.

Feature Engineering:

Consider additional feature engineering, such as sentiment analysis or named entity recognition, depending on your dataset and approach.

Save Preprocessed Data:

Save the preprocessed dataset to simplify future phases of the project.

.