Course : Artificial Inteligence

Project:Fake News Detection Using Natural Process Language

Document : Phase 2 Submit

Problem statement:

We consume news through several mediums throughout the day in our daily routine, but sometimes it becomes difficult to decide which one is fake and which one is authentic. Our job is to create a model which predicts whether a given news is real or fake.

1. **Problem Statement*:** Define the problem statement and the objective of the project.

2. **Data Gathering*:** Collect data from various sources that are relevant to the problem statement.

3. **Data Preprocessing*:** Perform some operations on the collected data to make it suitable for analysis. This includes:

Tokenization*: Break down the text into smaller units called tokens.

Lower Case*: Convert all text to lower case to avoid case sensitivity issues.

Stopwords*: Remove common words that do not add much meaning to the text, such as "the", "and", "a", etc.

Lemmatization / Stemming*: Reduce words to their base form to avoid redundancy and improve analysis accuracy.

4. **Vectorization (Convert Text data into the Vector)*:** Convert text data into a numerical format that can be used for analysis. This includes:

- *Bag Of Words (CountVectorizer)*: Create a matrix of word counts for each document in the dataset.

- *TF-IDF*: Assign weights to each word based on its frequency in the document and across all documents in the dataset.

5. **Model Building***: Build a machine learning model using the preprocessed and vectorized data. This includes:

- *Model Object Initialization*: Choose an appropriate machine learning algorithm and initialize a model object.

- *Train and Test the Model*: Split the dataset into training and testing sets, fit the model on the training set, and evaluate its performance on the testing set.

6. **Model Evaluation*:** Evaluate the performance of the machine learning model using various metrics. This includes:

- *Accuracy Score*: Calculate the proportion of correctly classified instances in the testing set.

- *Confusion Matrix*: Create a table that summarizes the number of true positives, true negatives, false positives, and false negatives in the testing set.

- *Classification Report*: Generate a report that summarizes various classification metrics such as precision, recall, F1-score, etc.

7. *Model Deployment*: Deploy the machine learning model in a production environment so that it can be used for real-world applications.

8. **Prediction of Client Data*:** Use the deployed machine learning model to make predictions on new data.

Teck Stack :

1. PYTHON
2. NLP
3. MACHINE LEARNING ALGORITHMS