Checkpoint 1: Data Preprocessing for Fake Review Detection

What is Data Preprocessing?

Data preprocessing is the foundational step in building any machine learning model. It involves cleaning, transforming, and organizing raw data into a format that can be efficiently used for training. This step ensures that the model receives high-quality input data, which is essential for achieving accurate predictions.

Basic Components of Preprocessing:

1. **Data Cleaning:**
    - Handling missing values by filling or removing incomplete data entries.
    - Removing duplicates and irrelevant data (e.g., spam or unrelated reviews).
2. **Text Normalization:**
    - Converting all text to lowercase to maintain uniformity.
    - Removing punctuation, special characters, and numbers where not required.
3. **Tokenization:**
    - Breaking down sentences into individual words or tokens for easier analysis.
4. **Stopword Removal:**
    - Eliminating common words (e.g., "and," "the") that do not add significant meaning.
5. **Stemming/Lemmatization:**
    - Reducing words to their root or base form (e.g., "running" → "run").

6. **Vectorization:**
    - Converting text data into numerical formats (e.g., TF-IDF or word embeddings) suitable for machine learning algorithms.

Tasks:

1. Load the raw dataset provided and explore its structure.
2. Perform data cleaning by handling missing values, duplicates, and irrelevant entries.
3. Normalize the text data by converting to lowercase and removing punctuation, special characters, and numbers.
4. Tokenize the reviews into individual words.
5. Remove stopwords to focus on meaningful terms.
6. Apply stemming or lemmatization to standardize word forms.
7. Convert the preprocessed text into numerical vectors using a suitable method (e.g., Bag-of-Words, TF-IDF, or embeddings).
8. Save the preprocessed dataset in a CSV file for further use.
9. Upload the code and the preprocessed dataset to GitHub with the repository titled **“Project_WoC_7.0_Fake_Review_Detection”** and inside it create folder for “**checkpoint 1”**.

Deadline:

10th January, 2024, 11:59 PM

Deliverables:

- GitHub repository with:
    - Preprocessing code.
    - A README file describing the preprocessing steps and results.
    - The preprocessed dataset in CSV format.