**Preprocessing Steps for NLP Project Dataset**

**1. Loading the Dataset**

We began by loading the dataset from a CSV file into a pandas DataFrame. This dataset included various columns such as text and label.

**2. Understanding the Dataset**

To get a good understanding of the dataset's structure, we displayed the first few rows. This step helped us identify the columns present and their respective data types.

**3. Removing Unnecessary Columns**

We removed any columns that were not needed for our analysis. In this case, we dropped the link column.

**4. Text Cleaning**

We performed extensive text cleaning on the statement column. This cleaning process involved:

* Converting text to lowercase.
* Removing line breaks and punctuation.
* Removing stop words using NLTK's stopwords list.
* Removing numbers.
* Lemmatizing the text using spaCy for normalization.

**5. Tokenization**

Next, we tokenized the cleaned text. Tokenization is the process of splitting text into individual words or tokens. We used NLTK's word\_tokenize method for this purpose.

**6. TF-IDF Vectorization**

We converted the tokenized text into numerical features using TF-IDF Vectorization. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (corpus).

**7. Mapping Veracity Labels**

The veracity column contained categorical data that needed to be standardized. We mapped the original categories to simplified versions, such as mapping 'True' and 'Mostly True' to 'True', 'False' and similar categories to 'False', and retaining 'Half-True' as it was.

**8. Label Encoding**

We encoded the cleaned veracity column into numerical values using LabelEncoder from sklearn. This step converted categorical labels into a format that can be used by machine learning algorithms.

**9. Saving the Preprocessed Data**

Finally, we saved the preprocessed DataFrame to a CSV file and downloaded it. This allowed us to easily share the preprocessed dataset.