**Dataset Creation and Annotation Process for Computational Exercise**

**Title:** Propaganda vs. Non-Propaganda Text Classification Dataset

**Objective:** The primary goal of this dataset is to facilitate the development and evaluation of machine learning models for the classification of textual content into 'Propaganda' and 'Non-Propaganda' categories.

**Data Collection:**

* Sources: The data was compiled from a variety of sources, including news articles, political speeches, and social media posts, to ensure diversity and representativeness.

**Annotation Process:**

* Annotation Team: The dataset was annotated by a team of 4 individuals with expertise in political science, journalism, and natural language processing.
* Guidelines: Annotators followed a detailed guideline that defines 'Propaganda' and 'Non-Propaganda' based on linguistic features, context, and intent. The guideline aimed to minimize subjectivity and ensure consistency across annotations.
* Review Process: Each entry was independently annotated by at least two team members. Discrepancies were resolved through discussion or third-party arbitration to ensure annotation quality.

**Data Format:**

* Each entry in the dataset consists of two fields:
  + Text: The textual content to be classified.
  + Label: The assigned label ('Propaganda' or 'Non-Propaganda').

**Ethical Considerations:**

* The dataset was curated with a focus on ethical considerations, including the avoidance of bias, respect for privacy, and the inclusion of a diverse range of viewpoints.

**Intended Use:**

* This dataset is intended for academic and research purposes, specifically for training and evaluating natural language processing models in the domain of propaganda detection.

**Limitations:**

We acknowledge potential limitations in the dataset, including possible biases and the inherent subjectivity in categorizing complex political and social narratives.

**Exploratory Data Analysis**

**Dataset Overview:**

* Total Records: 80 entries.
* Total Propaganda: 44 entries are labelled as 'Propaganda'.
* Total Non-Propaganda: 34 entries are labelled as 'Non-Propaganda'.
* Missing Values: There is 1 missing value in the dataset.

**Class Balance:**

* The dataset is relatively balanced but slightly skewed towards 'Propaganda':
* Propaganda: Approximately 55.17% of the entries.
* Non-Propaganda: Around 43.00% of the entries.

The data exploration provides several insights:

* **Label Counts:** The dataset contains two categories: 'Propaganda' and 'Non-Propaganda'. The bar chart illustrates the count of entries in each category. This distribution can help understand if the dataset is balanced or if there's a skew towards one label.
* **Average Text Length:**The analysis of the average text length for each category shows the mean length of text in the 'Propaganda' and 'Non-Propaganda' categories. This can indicate if propaganda texts tend to be longer or shorter than non-propaganda texts, which might be a characteristic feature.
* **Word Frequency Analysis:**
  + Top 10 most common words in each category:
    - In 'Propaganda' texts, the most common words include 'the', 'to', 'of', 'are', 'is', 'and', 'in', 'a', 'by'.
    - In 'Non-Propaganda' texts, the top words are 'to', 'the', 'in', 'and', 'of', 'a', 'for', 'economic'.
  + This word frequency analysis reveals the commonly used words in each type of text. While there is some overlap (common words like 'the', 'to', 'of', 'and', 'in' appear in both), the context in which these words are used might differ, contributing to the distinction between propaganda and non-propaganda texts.

**Insights**:

* Balanced Dataset: The dataset has a good balance between the two classes, which is beneficial for training a classification model. However, the slight skew towards 'Propaganda' should be kept in mind.
* Consistent Text Length: The texts are relatively short and have a moderate variation in length. This consistency in length could be beneficial for certain types of NLP models, as they might not need to handle very large discrepancies in input size.
* Text Analysis for Feature Engineering: Given the moderate variation in text length, exploring features like word count, sentence complexity, and use of specific keywords or phrases might provide additional insights and improve model performance.
* Contextual Understanding: Since propaganda detection involves understanding nuances and context, advanced NLP models like RoBERTa, which can capture contextual information, might perform better than simpler models.

**Research paper and Image reference link :**

* <https://hal.science/hal-03417019/document>
* <https://peerj.com/articles/cs-1248/>
* <https://arxiv.org/pdf/2305.14534v1.pdf>
* <https://arxiv.org/pdf/2310.06422v1.pdf>
* <https://aclanthology.org/2020.semeval-1.229.pdf>
* <https://aclanthology.org/D19-5014/>
* <https://aclanthology.org/2022.wanlp-1.61.pdf>
* <https://arxiv.org/pdf/2310.06422v1.pdf>

**Image Reference Link :**

* <https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.cnn.com%2F2022%2F03%2F05%2Fpolitics%2Ffact-check-fake-cnn-ukraine%2Findex.html&psig=AOvVaw2xjo9TG0Rkoc1Pz8nmSH-a&ust=1701482628668000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCOC6kO-S7YIDFQAAAAAdAAAAABAE>
* <https://en.detector.media/post/trump-musk-and-pinocchio-how-russian-agitational-propaganda-uses-fakes-to-search-for-non-standard-connections-with-the-usa?fbclid=IwAR3ipz8hpNEPpkekJmQANQZ_QhzcJjCeiPJP2y0IBEuRqcZSZSs2_YmBz60>
* <https://propaganda.qcri.org/>