# Data Acquisition and Pre-processing Report

For our fake news detection project, we will be utilizing the **LIAR dataset**, which is hosted on Hugging Face and can be accessed here. This dataset contains **12.8K human-labelled short statements** from Politifact.com, making it an ideal resource for building a fake news detection system. The data has been labelled by Politifact.com editors, providing a comprehensive assessment of the truthfulness of each statement.

**Metadata**

- **Size**: The dataset consists of 12,836 labelled statements.
- **Features/Columns**: ["id", "label", "statement", "subject", "speaker", "job title", "state_info", "party_affiliation", "barely_true_counts", "false_counts", "half_true_counts", "mostly_true_counts", "pants_onfire_counts", "los", "justification"]

**Pre-processing Plan**

We will carry out several pre-processing steps. Some important once are listed below:

1. **Handling Missing Values**: We will inspect the dataset for any missing or incomplete data. Missing values will either be filled in with appropriate techniques (e.g., filling missing job titles) or the corresponding rows will be removed if necessary.
2. **Text Cleaning**: Since the primary feature is the "statement", we will clean the text by removing special characters, converting to lowercase, and performing tokenization.
3. **Label Encoding**: The "label" column, which contains categorical truthfulness labels, will be encoded into numeric form for use in machine learning algorithms.
4. **Feature Engineering**: We will experiment with using additional metadata, such as "subject", "party_affiliation", and "justification" as input features, which could provide valuable contextual information for the classification task.
5. **Balancing the Dataset**: While the dataset is fairly balanced, we will check the distribution of labels and apply techniques like undersampling or oversampling.

**Usage of data**

Our main focus will be on using the "statement" as the input feature and the "label" as the target variable. Initially, we will begin by treating this as a multi-class classification problem to categorize statements into one of the six truthfulness labels. We will also explore whether additional columns like "subject", "party", and "justification" can enhance the model's performance through feature engineering.

**Initial Thoughts on Methods**

For our initial approach, we plan to start with traditional machine learning algorithms such as **Logistic Regression**, **Support Vector Machines (SVM)**, and **Random Forests** to establish baseline performance. Following that, we will move to more advanced methods like **Transformer-based models (BERT)** and **RNN/LSTM networks** for processing the text in the "statement" field. Since fake news detection is a complex task requiring an understanding of linguistic nuances, deep learning methods such as **fine-tuned BERT models** may prove particularly effective.