# Project Title: Fake News Detection Using NLP and Machine Learning

## Objectives

-Project scope and methodology.  
- Obtain and clean the LIAR dataset for modeling.

LIAR dataset remains a common benchmark for short-form fake news detection.

## 1. Scope Definition

- Problem: Classify short political statements as fake or real.  
- Dataset: LIAR dataset (downloaded).  
- Models to Explore: TF-IDF + classical ML, Word2Vec + LSTM, BERT.  
- Optional Output: Flask-based prediction tool.

## 2. Dataset Collection – LIAR

- Source: William Yang Wang's LIAR dataset  
- Files: train.tsv, valid.tsv, test.tsv (focus on train.tsv this week)  
- Dataset Summary:  
 - 12,836 short political statements.  
 - Labels: pants-fire, false, barely-true, half-true, mostly-true, true.  
 - Metadata: speaker, party, context, subject, etc.

A diagram of a table

Description automatically generated with medium confidence

## 3. Data Cleaning

- Loaded train.tsv using Pandas.  
- Removed missing/null entries.  
- Standardized casing and punctuation.  
- Cleaned statement field for further NLP (stopword removal, lemmatization scheduled for Week 2).

## 4. Status Summary

|  |  |
| --- | --- |
| Task | Status |
| Scope Defined | ✅ |
| LIAR Dataset Collected | ✅ |
| Data Cleaning Started | ✅ (basic) |
|  |  |