PROBLEM STATEMENT

Project: Automatic Detection of Social Bias in Text

Objective:

This project aims to build an NLP-based system that automatically detects and highlights social biases (e.g., gender, race, age) in written content such as media articles, social media posts, and online comments. The system will flag potentially biased language and provide feedback to help users understand and mitigate their use of biased terms or stereotypes.

Approach:

- **Preprocessing**: The text is cleaned and prepared for analysis by removing irrelevant characters, tokenizing, and normalizing.
- **Embedding Generation**: Word embeddings are generated using advanced NLP techniques:
    - **ELMo**: Provides word-level, context-sensitive embeddings that capture bias in specific terms within varying contexts (e.g., "doctor" in relation to male or female pronouns).
    - **BERT**: Captures the broader context and sentence-level bias through its transformer-based architecture.
- **Model Training**: A classification model is trained using labeled data to detect various forms of bias (gender, racial, age-related, etc.) in text. The model learns patterns in language associated with bias and non-bias.
- **Bias Detection**: The trained model analyzes new input text, highlighting sections that exhibit potential bias.

Tools:

- **Python** for development and implementation.
- **Scikit-learn** for classification models and preprocessing tasks.
- **ELMo** for word-level context-sensitive embeddings.
- **BERT** for deep contextual language understanding.

Extensions:

- **Feedback System**: A user-facing interface provides real-time feedback, suggesting alternative, unbiased phrases when biased content is detected.
- **Bias Explanations**: Include an explainability feature using BERT's attention mechanism to show why the text was flagged as biased, helping users learn and improve their language usage.