AI ASSIGNMENT 01
CLARITY - Text-Based Political Interview Analysis

Muhammad Hanan Zia (515271)
Muhammad Umar Tahir (507955)
Muhammad Ibrahim (500927)

November 16, 2025

1 Introduction and Motivation

Political interviews are a rich source of insights into communication strategies, where speakers may clarify or evade questions. Systematic analysis of these texts can reveal patterns in language use and response behavior.

The CLARITY project develops a text-based exploratory data analysis framework for political interviews, using NLP techniques to examine token lengths, n-grams, vocabulary, language, sentiment, and label distributions for clarity and evasion.

The project is motivated by three goals:

Data Understanding: Explore dataset structure, label distribution, and potential issues like noise or imbalance.

Text Analysis Skills: Apply NLP methods such as tokenization, n-gram analysis, vocabulary assessment, and sentiment detection.

Interpretability and Visualization: Generate plots and metrics to extract meaningful, interpretable insights that guide further analysis or modeling.

This framework provides a foundation for systematically studying political discourse and response patterns.

2 Dataset Overview

The CLARITY dataset contains political interview responses with clarity and evasion labels, split into train and test sets.

Table 1: CLARITY Dataset Information

Attribute	Description
Dataset Name	CLARITY Political Interview Dataset
Dataset Source (Hugging Face)	https://huggingface.co/datasets/ailsntua/QEvasion
Languages Covered	English
Files in Train Set	3,448 text entries
Files in Test Set	308 text entries
Modalities	Text (interview questions and answers)
Columns / Features	`interview_question`, `interview_answer`, `title`, `date`, `president`, `url`, `clarity_label`, `evasion_label`, annotator columns, flags (`inaudible`, `multiple_questions`, `affirmative_questions`)
Average Text Length	$\approx$ 294 tokens (mean of `interview_answer`)
Total Labels	Positive and negative clarity/evasion labels for each interview answer
File Format	.csv (train.csv, test.csv)
Notes	Suitable for text-based EDA and NLP tasks (tokenization, n-grams, sentiment analysis, label distributions)

2.1 Train/Test Counts

Table 2: Train and Test Dataset Counts

Dataset Split	Number of Entries
Train	3,448
Test	308

2.2 Train Dataset Feature Overview

Table 3: Train Dataset Column Overview ($N = 3,448$)

Column	Type	Missing	Unique
title	object	0	175
date	object	0	287
president	object	0	4
url	object	0	287
question_order	int64	0	73
interview_question	object	0	2,061
interview_answer	object	0	2,004
gpt3_5_summary	object	0	2,390
gpt3_5_prediction	object	0	2,111
question	object	0	3,386
annotator_id	int64	0	3
annotator1	float64	3,448	0
annotator2	float64	3,448	0
annotator3	float64	3,448	0
inaudible	bool	0	2
multiple_questions	bool	0	2
affirmative_questions	bool	0	2
index	int64	3,448	0
clarity_label	object	0	3
evasion_label	object	0	9

2.3 Train Dataset Token Statistics

Table 4: Text Columns Token Length Statistics (Train Dataset)

Column	Mean	Median	Min	Max
question_order	1.00	1	1	1
interview_question	61.51	50	3	780
interview_answer	293.57	207	1	2,117
question	14.46	12	1	80
multiple_questions	1.00	1	1	1
affirmative_questions	1.00	1	1	1

2.4 Label Distributions

Table 5: Clarity Label Distribution

Clarity Label	Count
Ambivalent	2,040
Clear Reply	1,052
Clear Non-Reply	356

Table 6: Evasion Label Distribution

Evasion Label	Count
Explicit	1,052
Dodging	706
Implicit	488
General	386
Deflection	381
Declining to answer	145
Claims ignorance	119
Clarification	92
Partial/half-answer	79

3 Key Features for EDA

- **Text Analysis:** Token lengths, vocabulary, sentence statistics, n-grams, sentiment.
- **Labels:** Study distributions and correlations for clarity and evasion.
- **Challenges:** Noise, class imbalance, missing data, variation in text length.

4 Problem Formulation and Evaluation Metrics

Problem Definition

Predict clarity and evasion labels for each interview response.

Input / Output

- **Input:** Interview question and answer text.
- **Output:** Probability / label for clarity and evasion.

Evaluation Metrics

- Label Distribution Accuracy
- Macro F1 Score
- Token / EDA metrics (tokenization, n-grams, sentiment)

5 Task Division Among Group Members

Table 7: Task Division for CLARITY AI Project

Member No.	Name	Assigned Tasks and Libraries Used
1	**Muhammad Hanan Zia**	Python implementation: data loading, preprocessing, tokenization, and automation of EDA scripts using **Pandas, NumPy, Matplotlib, Seaborn, NLTK, SpaCy, WordCloud, scikit-learn**. Generated plots for **token length distributions, sentence length statistics, word clouds, vocabulary size, n-gram frequency, missing data and flag distributions, and correlation heatmaps.**
2	**Muhammad Umar Tahir**	Visualization and analysis using **Matplotlib, Seaborn, Plotly**. Produced **clarity and evasion label distributions, multi-feature comparison plots, trend analysis of question/answer lengths, annotator agreement visualizations, and token length histograms by label.**
3	**Muhammad Ibrahim**	Validation and documentation using **Pandas, Matplotlib, Seaborn**. Reviewed code and plots for **accuracy and consistency, verified train/test splits, checked feature extraction and EDA outputs, documented findings, and finalized report tables, figures, and captions.**

6 GitHub Repository

https://github.com/HananZia/CLARITY.git