CLARITY / "I Never Said That" — Dataset Exploration & Analysis

Fatima Malik[*], **Shuja Naveed**[*], **Sumera Bibi**[*] **and Mehwish Fatima**[†]
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology, Islamabad, Pakistan
{fmalik.bscs24seecs, snaveed.bscs24seecs, sbibi.bscs24seecs,
mehwish.fatima,}@seecs.edu.pk

Abstract

This assignment expands a comprehensive exploratory data analysis (EDA) and implementing baseline classification pipelines for clarity and evasion detection. The dataset contains 3,448 annotated question answer pairs from presidential interviews, labeled for clarity (clear, unclear, evasive) and evasion techniques. Assignment 02 focuses on dataset understanding, visualization, label distribution analysis, token statistics, text normalization, and baseline implementations including TF–IDF + Linear SVM and BiLSTM models. This report summarizes dataset patterns, class imbalances, linguistic behavior, and baseline model performance to guide later improvements.

1 Dataset Overview

1.1 Dataset Description

The dataset consists of political interview transcripts annotated for response clarity and evasion. It includes 20 columns covering text fields, metadata, and multi-annotator labels. This assignment uses both the full training set and a 50-sample exploration subset.

- **Source:** CLARITY / "I Never Said That"
- **Main text fields:** question, interview_answer
- **Labels:** clarity_label (all instances), evasion_label (subset)
- **Metadata:** title, date, president, inaudible, multiple_questions, annotators
- **Files used:** train.parquet, test.parquet, train_sample50.csv

Dataset Summary Table

1.2 Data Structure

Columns include title, date, president, url, question_order, interview_question, interview_answer, annotator labels, clarity_label, and evasion_label.

Metric	Value
Number of rows (train)	3448
Number of rows (test)	308
Number of columns	20
Question missing values	0
Answer missing values	0
Inaudible count	45
Multiple_questions count	86

Table 1: Dataset summary statistics

1.3 Sample Records

A 50-row sample (train_sample50.csv) was used for initial exploration and visualization validation.

2 Exploratory Text Analysis

2.1 Word Count Analysis

Word-count histograms were generated for both questions and answers. Observations include:

- Questions tend to be concise and topic-focused.
- Answers vary significantly in length; evasive answers are often longer due to diversions.
- Some answers are extremely short or marked as inaudible, affecting clarity classification.

2.2 Token & N-gram Analysis

Tokenization and n-gram extraction were performed for both questions and answers. Findings:

- High-frequency question terms revolve around governance, foreign policy, and accountability.

[*]Equal contribution.
[†]Corresponding author.

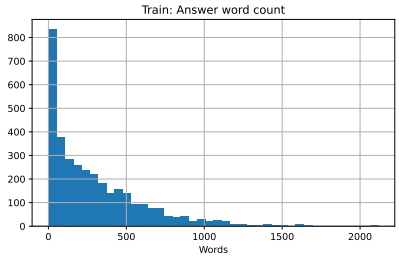

Figure 1: Distribution of word counts in interview answers.

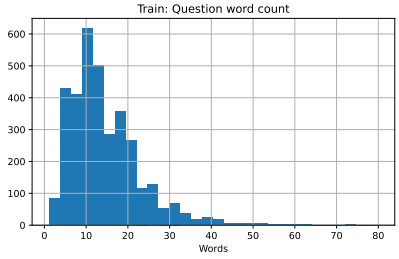

Figure 2: Distribution of word counts in questions.

- Common answer unigrams include discourse markers such as *well*, *look*, *let me*, often linked to evasion.
- Bigram analysis reveals turn-taking patterns reflecting political framing strategies.

Files generated:

- train_question_unigrams.csv
- train_answer_unigrams.csv
- train_question_bigrams.csv
- train_answer_bigrams.csv

3 Label Distribution & Data Challenges

3.1 Clarity Label Distribution

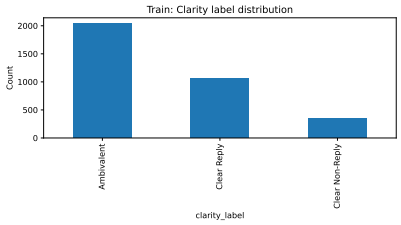

Figure 3: Distribution of clarity labels in the training set.

Clear responses dominate the dataset, followed by unclear responses. Evasive responses constitute the smallest proportion, creating a significant imbalance that hinders classifier performance on minority labels.

3.2 Evasion Label Distribution

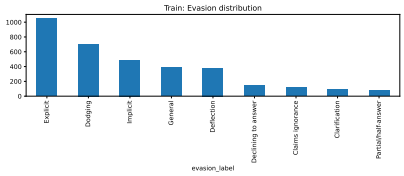

Figure 4: Distribution of evasion labels in the training set.

Evasion labels are extremely sparse, making accurate prediction challenging. Most answers are classified as non-evasive.

3.3 Annotated Examples

Examples illustrate clarity categories and typical linguistic patterns.

Label	Question	Answer
Clear	U.S.–China relations?	Response directly addresses geopolitical stance.
Ambivalent	China slowdown risks?	A partially addressing but vague reply.
Non-Reply	Putin's future actions?	Topic diverted toward past election interference.

3.4 Data Quality Issues & Challenges

- **Ambiguous labels:** Multi-annotator disagreements.
- **Incomplete answers:** Some clipped or missing due to transcription issues.
- **Inaudible segments:** Affect semantic clarity.
- **Class imbalance:** Significant skew toward clear and non-evasive categories.

4 Conclusion

It establishes the foundational understanding required for modeling political clarity and evasion. Dataset analysis reveals linguistic traits, structural issues, and class imbalance challenges. The implemented baselines—TF–IDF + SVM and BiLSTM provide initial performance metrics to refine in Assignment 03.