Assignment 1

Dimensional Stance Analysis (ParselQ)

Group Members

Muhammad Ahmad Amin (502217)
Hassan Jamal (519530)
Haniya Farhan (492237)
Syeda Frozish Batool (501165)

Department of Computer Science
Faculty of Computing
National University of Sciences and Technology, Islamabad

November 16, 2025

Contents

1 Introduction

This report presents the comprehensive Exploratory Data Analysis (EDA) for the project **Dimensional Stance Analysis (ParselQ)**, developed as part of Assignment 1. The task focuses on understanding stance variations in an environmental protection dataset under Track B, Subtask 1.

To complete the EDA efficiently, the work was divided equally among four group members. Each member was responsible for one major component as shown in the table below:

Task Division Table

Member	**Assigned Task**	**CMS ID**
Muhammad Ahmad Amin	Textual Token-Level + Valence - Arousal Label Analysis	502217
Hassan Jamal	Dataset Overview + Structure Validation	519530
Haniya Farhan	Aspect Frequency + Label Behavior Analysis	492237
Syeda Frozish Batool	Latex Report	501165

The remaining chapters provide detailed explanations and allocated plot spaces for each part.

2 Part 1: Dataset Structure, Validation, and Overview

2.1 Detailed Explanation

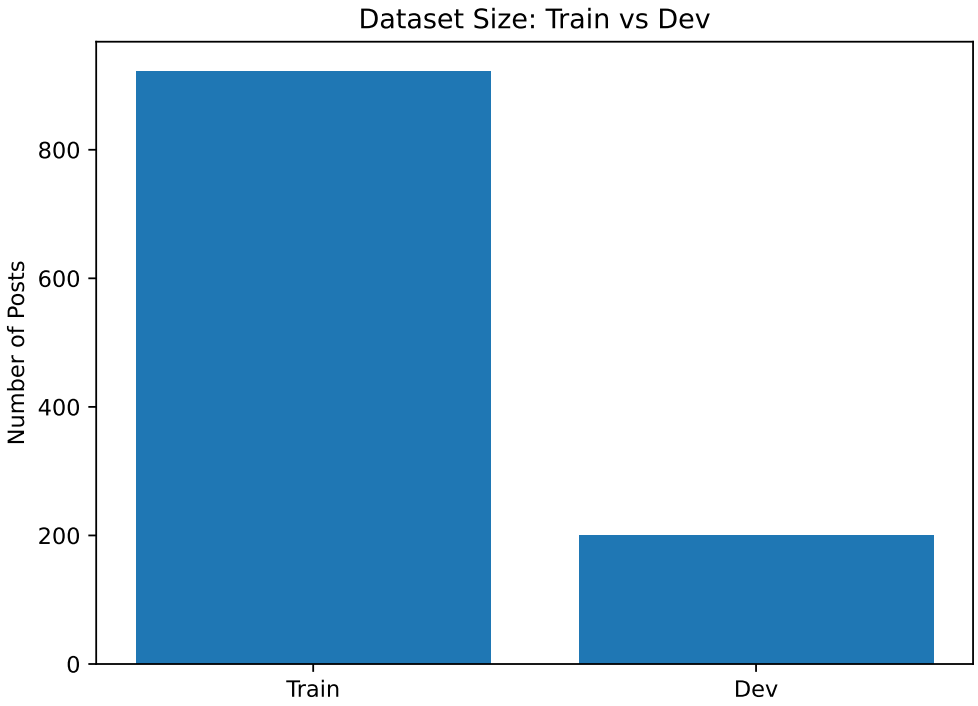
This component establishes the foundation of the entire EDA by examining the raw dataset structure. The dataset is provided in JSONL format, where each line represents a post containing textual content and a list of annotated stance aspects.

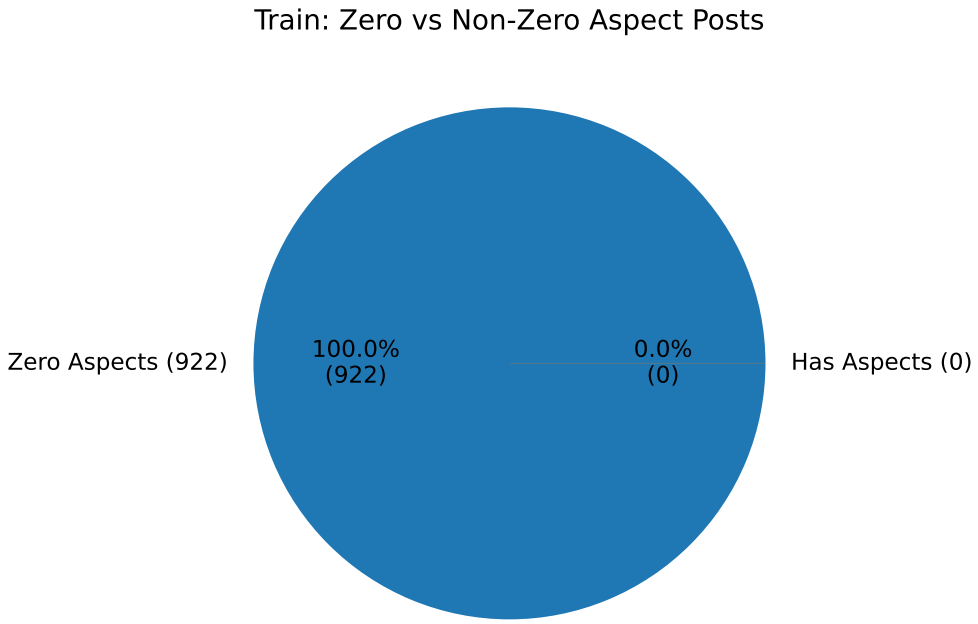
The key goals of this part include:

- Examining structural integrity of the dataset
- Identifying missing or malformed entries
- Computing essential statistics such as total aspects, average aspects per post, and missing fields
- Performing consistency checks for format, field types, and annotation validity
- Manually inspecting samples to ensure annotations align with text meaning

This structural analysis ensures that later EDA steps do not suffer from hidden dataset inconsistencies, which can corrupt model training.

2.2 Plots


Dataset Size: Train vs Dev
Number of Posts
0
200
400
600
800
Train
Dev


Train: Zero vs Non-Zero Aspect Posts
Zero Aspects (922)
100.0%
(922)
0.0%
(0)
Has Aspects (0)

3 Part 2: Textual Token Analysis

3.1 Detailed Explanation

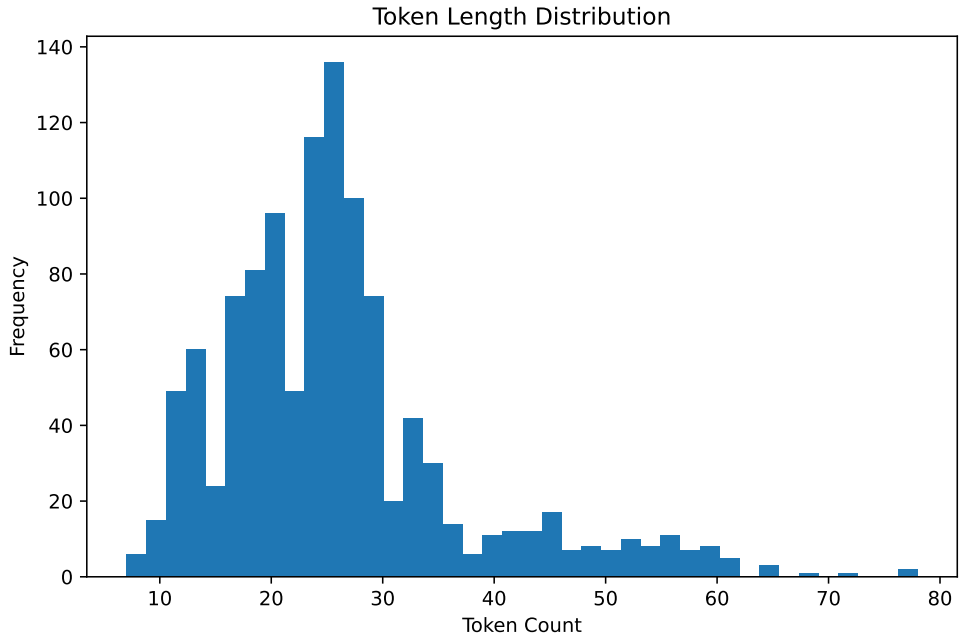
Linguistic analysis is essential for understanding the nature of textual content in stance detection tasks. This part examines the token-level structure of the posts using NLTK.

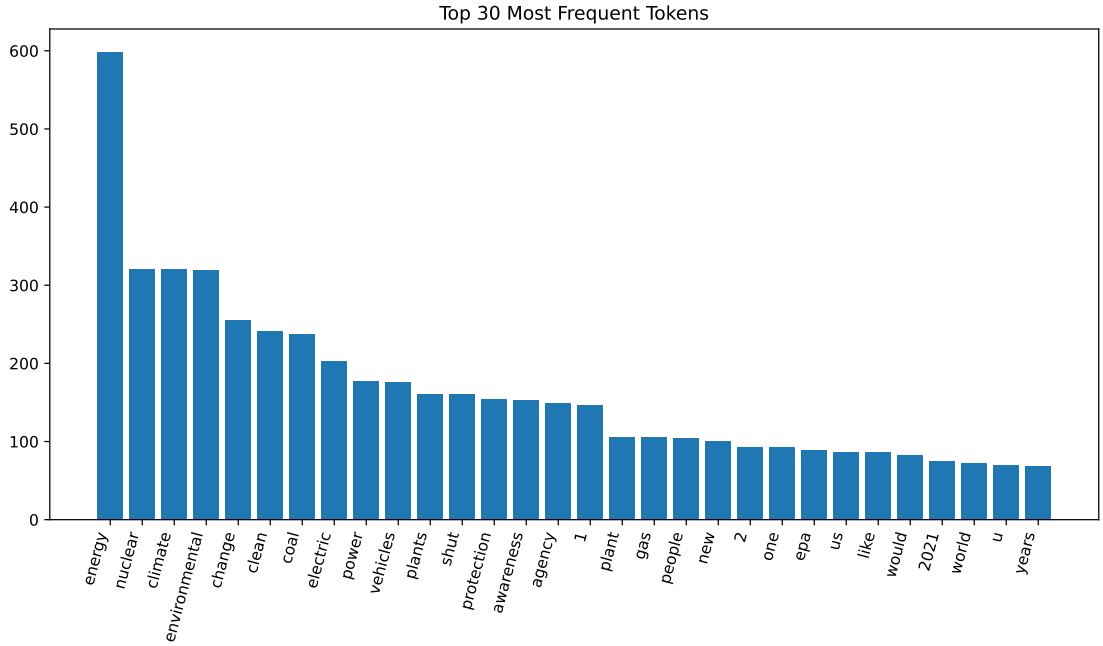
The analyses conducted include:

- **Token length distribution:** identifies short vs. long posts and outliers
- **Vocabulary richness:** measures lexical uniqueness and domain specificity
- **Stopword frequency:** reveals conversational vs. formal tendencies
- **N-gram extraction:** uncovers commonly co-occurring terms hinting at stance framing

These insights guide preprocessing decisions such as stemming, removal of stopwords, and handling of rare words.

3.2 Plots


Token Length Distribution
Frequency
Token Count
0
20
40
60
80
100
120
140
10
20
30
40
50
60
70
80


Top 30 Most Frequent Tokens
0
100
200
300
400
500
600
energy
nuclear
climate
environmental
change
clean
coal
electric
power
vehicles
plants
shut
protection
awareness
agency
1
plant
gas
people
new
2
one
epa
us
like
would
2021
world
u
years

4 Part 3: Aspect Distribution and Label Behavior

4.1 Detailed Explanation

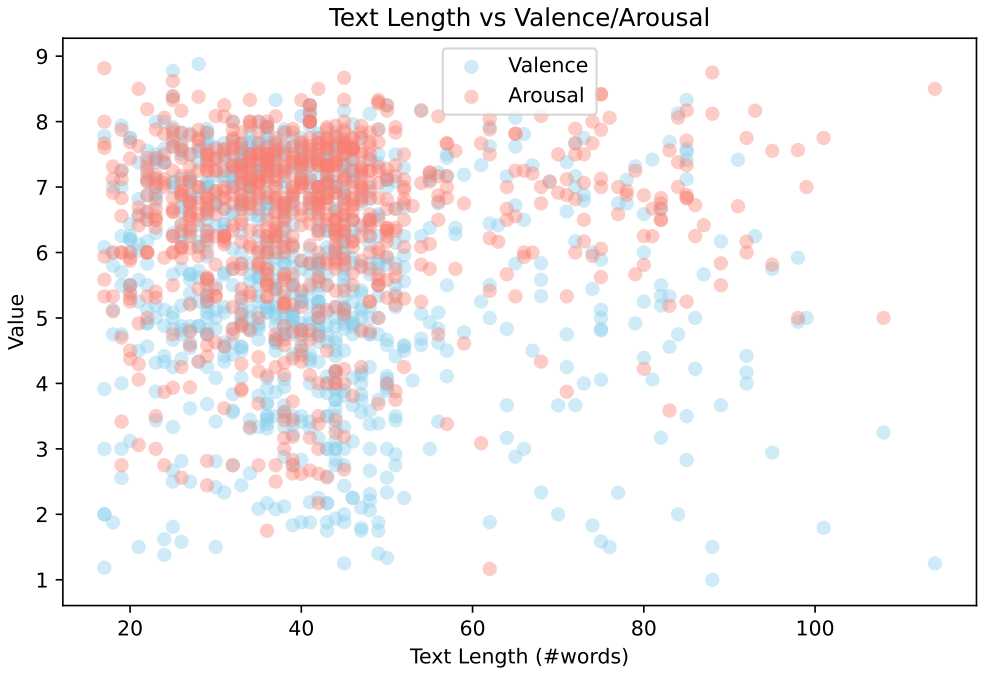
This segment dives into the stance aspects themselves — their frequency, distribution, and relationships. This analysis is crucial because label imbalance directly affects classification performance.

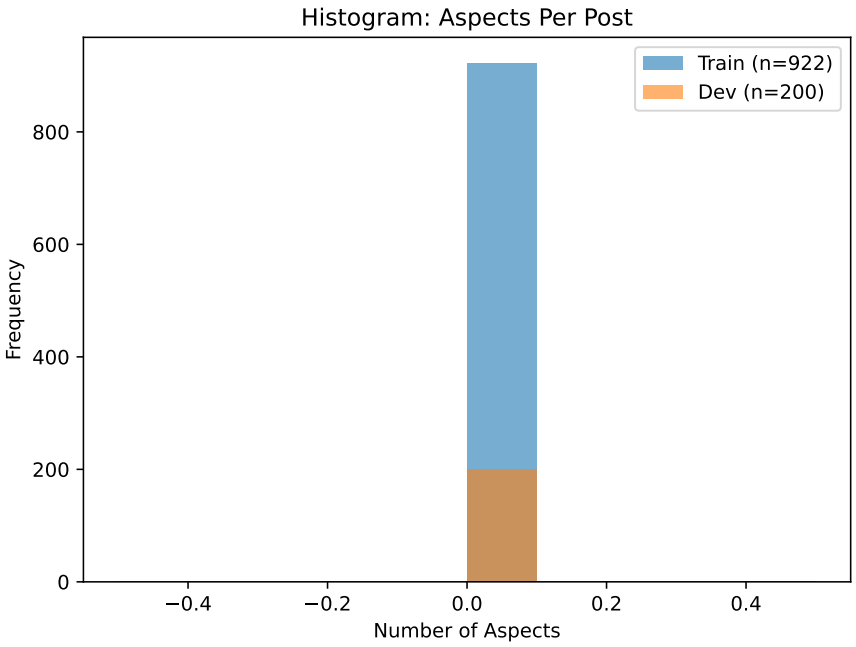
The following analyses were performed:

- **Aspect frequency distribution:** to identify dominant stance aspects
- **Skewness/imbalance detection:** to determine need for resampling or class weights
- **Aspect co-occurrence:** reveals relationships and frequently paired stance dimensions
- **Summary statistics:** aspects/post, variance, and distribution shape

Understanding label behavior ensures a fair and robust downstream ML model.

4.2 Plots


Text Length vs Valence/Arousal
Valence
Arousal
Value
Text Length (#words)
1
2
3
4
5
6
7
8
9
20
40
60
80
100


Histogram: Aspects Per Post
Train (n=922)
Dev (n=200)
Frequency
Number of Aspects
0
200
400
600
800
−0.4
−0.2
0.0
0.2
0.4

5 Part 4: Valence - Arousal Label Analysis

5.1 Detailed Explanation

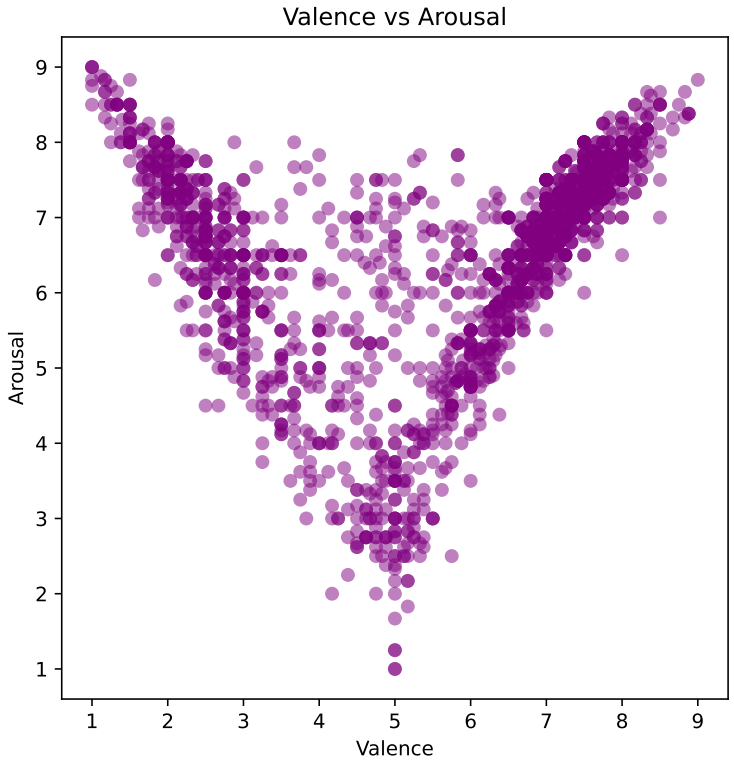
This part focuses on the valence and arousal (VA) labels, which are available **only in the training set**. Visualizing these labels helps understand the emotional distribution of the dataset.

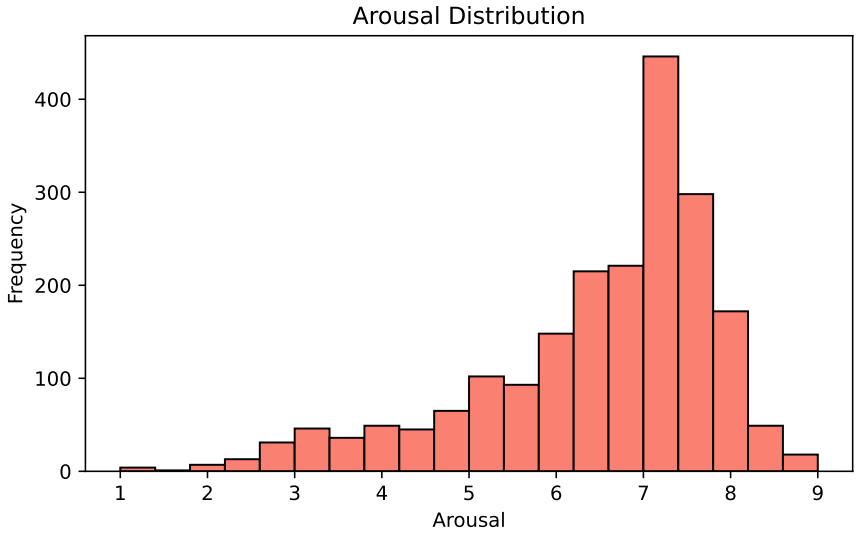
The insights include:

- Distribution of Valence scores
- Distribution of Arousal scores
- Relationship between Valence and Arousal (scatter plot)
- Correlation between Valence and Arousal

These observations highlight whether the VA labels are balanced or skewed, whether texts are generally positive or negative, and the prevalence of high-arousal posts. Such insights are crucial for designing models that effectively capture affective dimensions.

5.2 Plots


Valence vs Arousal
Arousal
Valence
1
2
3
4
5
6
7
8
9


Arousal Distribution
Frequency
Arousal
0
100
200
300
400
1
2
3
4
5
6
7
8
9

GitHub Repository

(https://github.com/muhammad-ahmad-amin/ParselQ.git)

Thank You!