Dataset Dimensional Analysis Report

Generated by Python Script

November 18, 2024

Analysis of reduced_emotion_test

Basic Information

- Number of samples: 2000
- Missing 'text' entries: 0
- Missing 'label' entries: 0
- Number of unique sentences: 2000
- Number of unique labels: 6
- Vocabulary size: 4796

Sentence Length (Words)

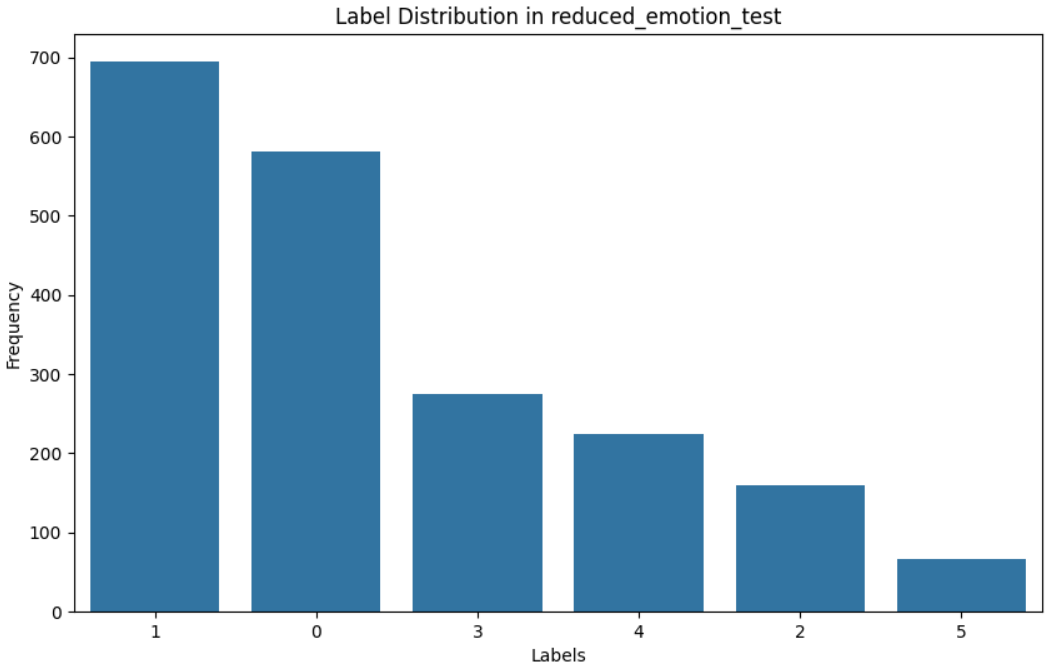
- Average: 19.15
- Standard deviation: 11.01
- Median: 17.0
- Max: 61
- Min: 3
- Quantiles (25%, 50%, 75%): 10.0, 17.0, 26.0

Stop Words Proportion

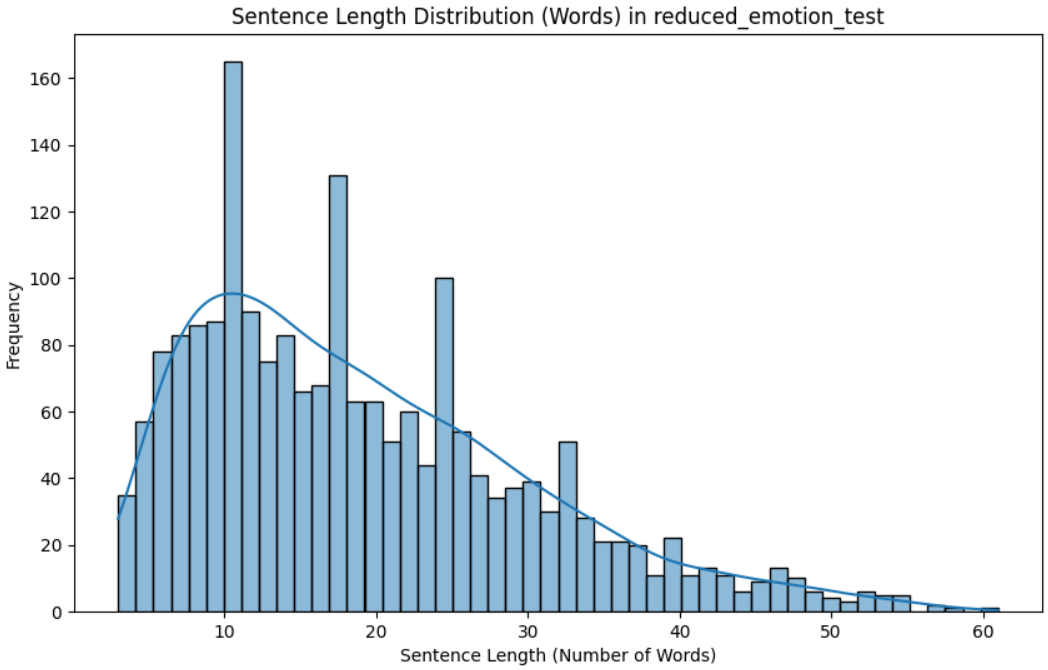
51.44%

Label Distribution

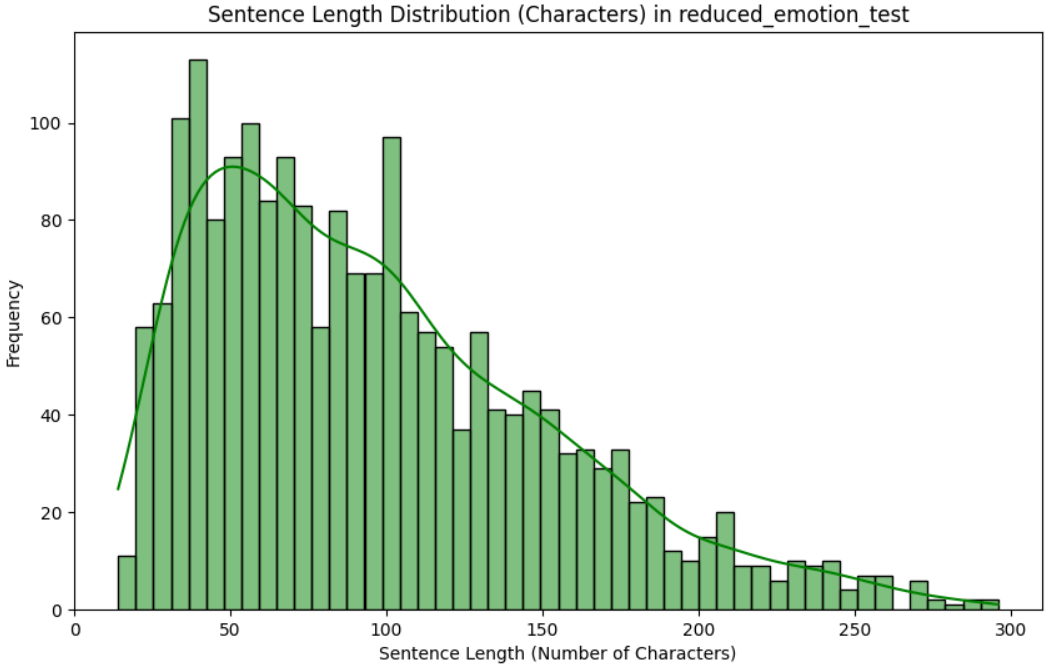
Label	Frequency
1	695
0	581
3	275
4	224
2	159
5	66



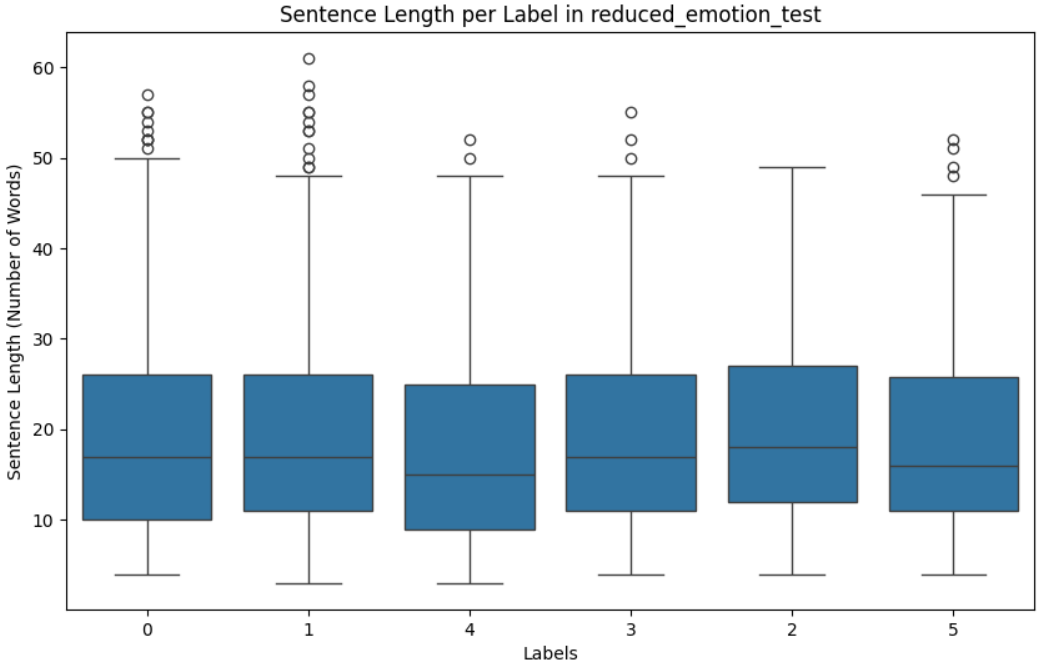
Sentence Length Distribution (Words)



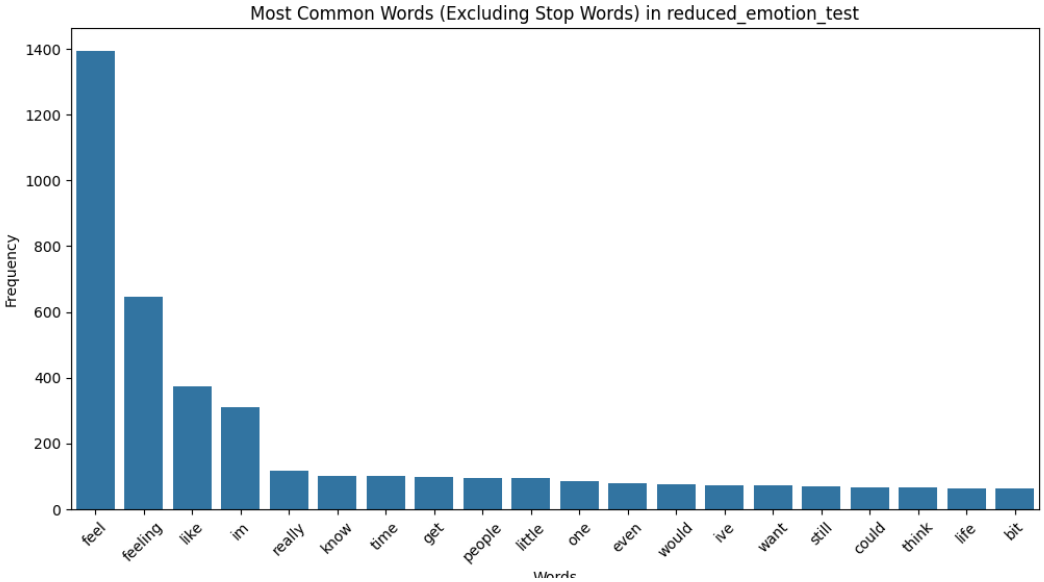
Sentence Length Distribution (Characters)



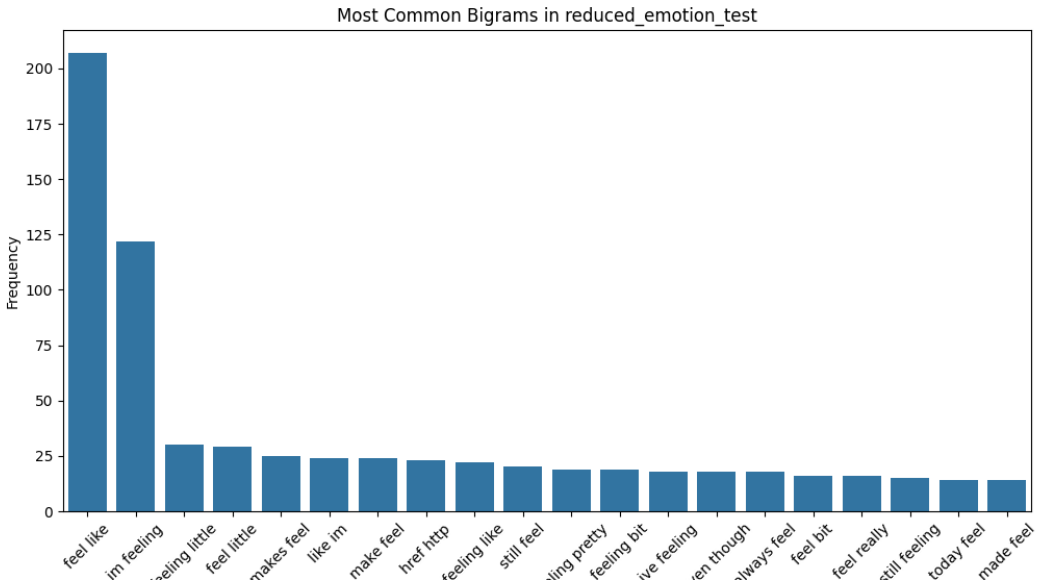
Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams



Analysis of reduced_emotion_train

Basic Information

- Number of samples: 6000
- Missing 'text' entries: 0
- Missing 'label' entries: 0

- Number of unique sentences: 5997
- Number of unique labels: 6
- Vocabulary size: 9034

Sentence Length (Words)

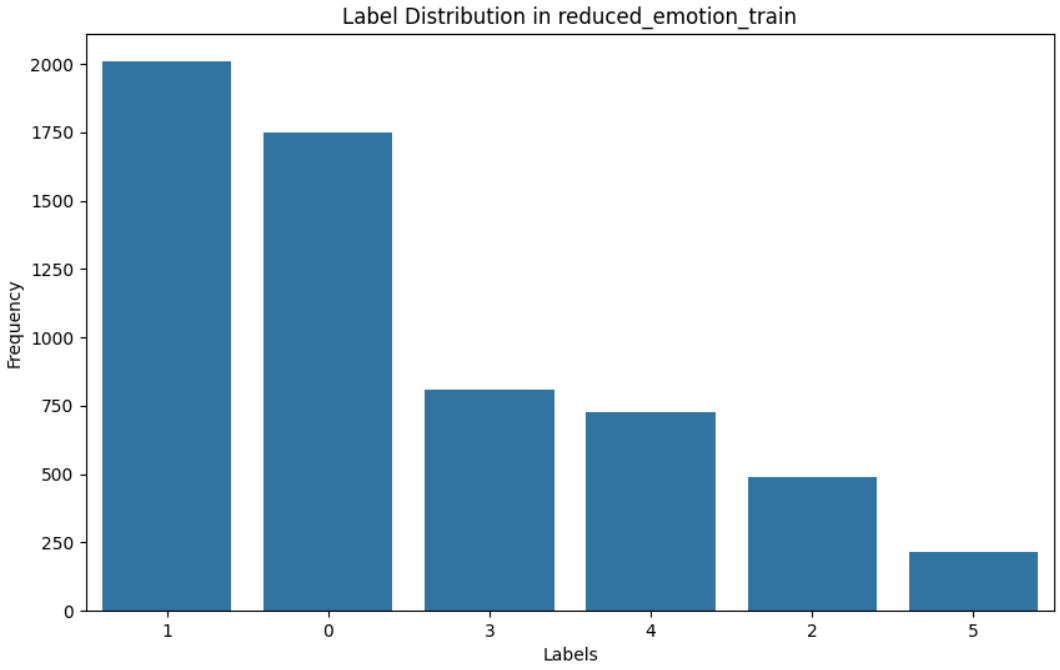
- Average: 19.07
- Standard deviation: 11.06
- Median: 17.0
- Max: 64
- Min: 2
- Quantiles (25%, 50%, 75%): 11.0, 17.0, 25.0

Stop Words Proportion

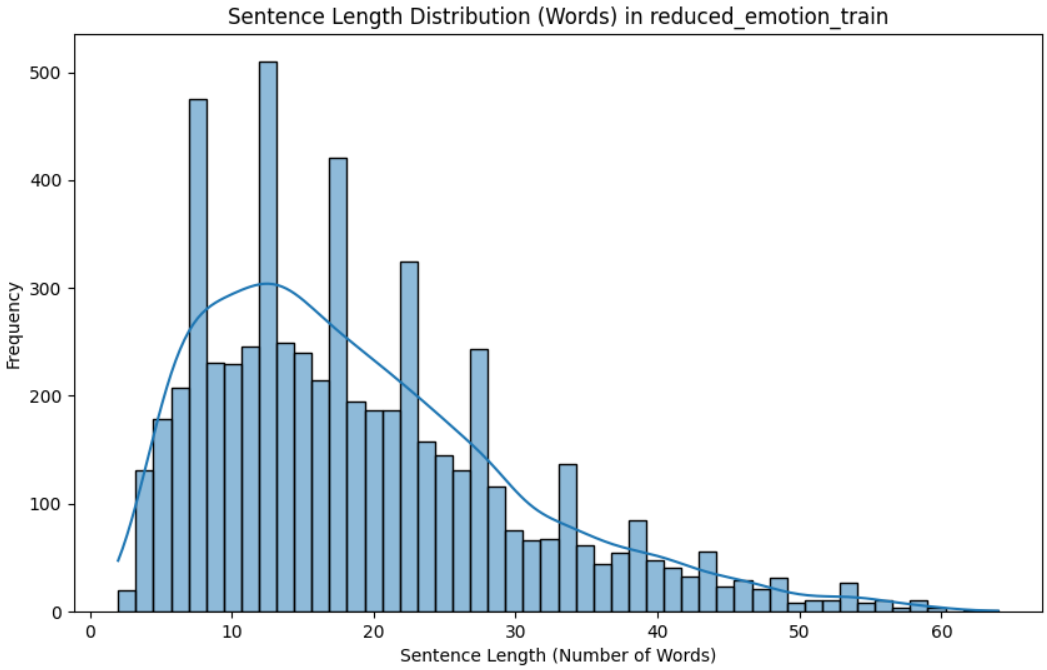
51.18%

Label Distribution

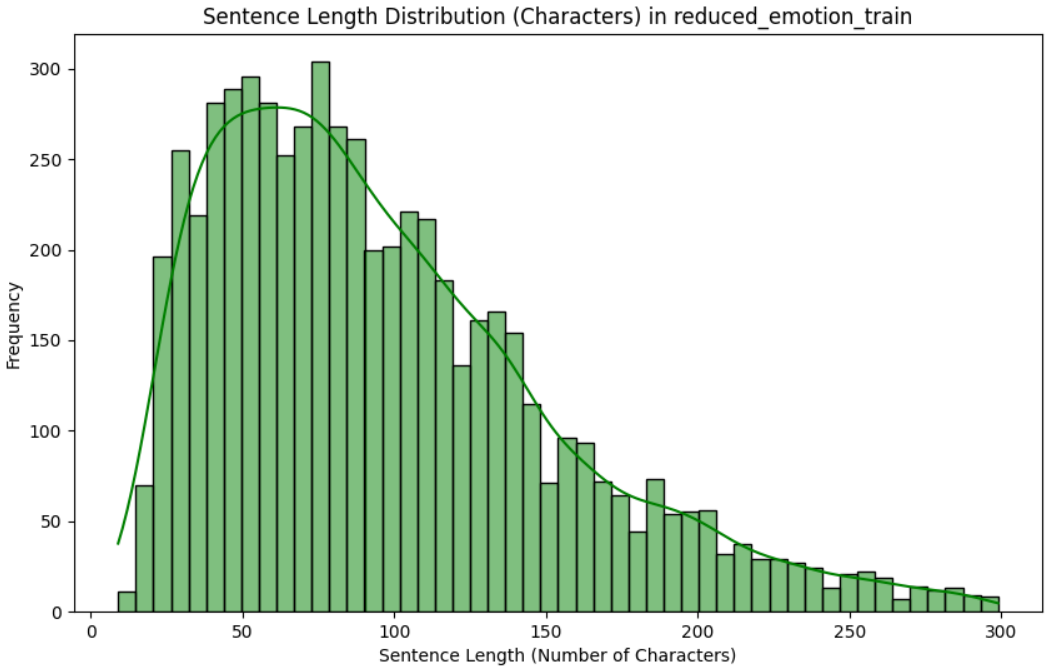
Label	Frequency
1	2010
0	1750
3	810
4	726
2	490
5	214



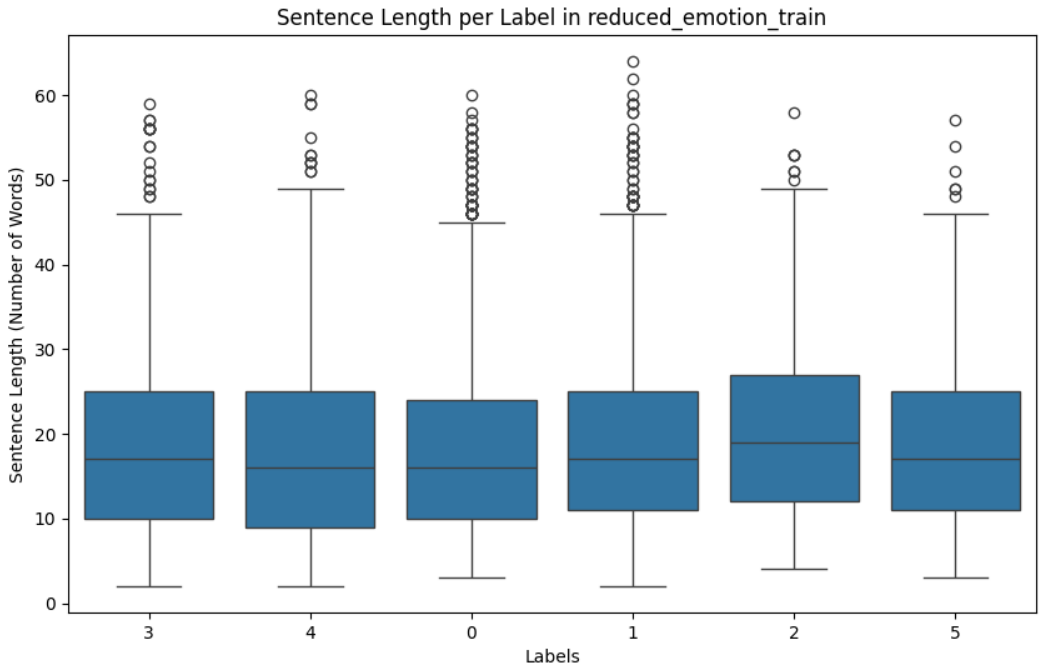
Sentence Length Distribution (Words)



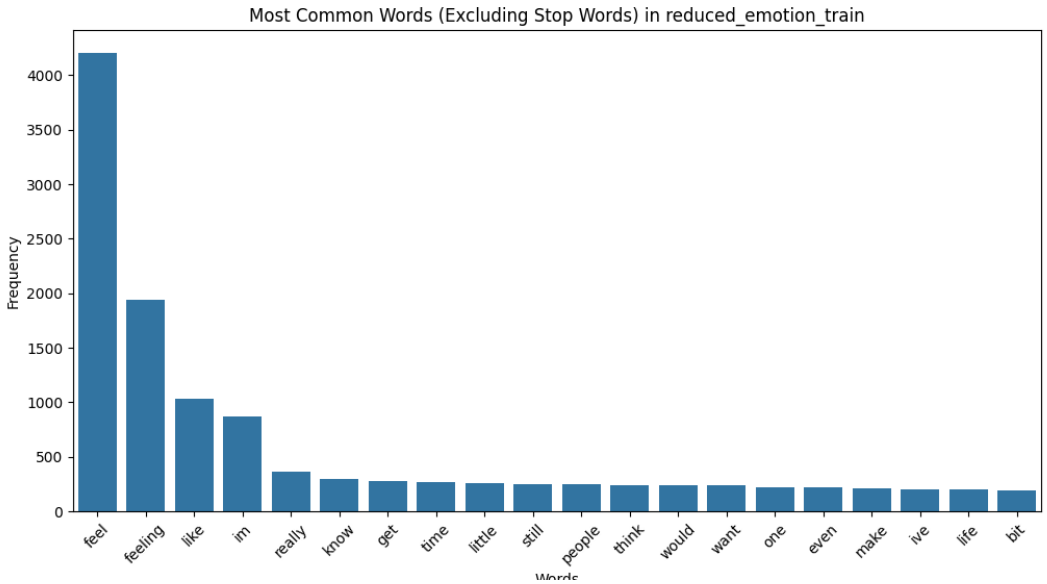
Sentence Length Distribution (Characters)



Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams