Dataset Dimensional Analysis Report

Generated by Python Script

October 23, 2024

Analysis of 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)

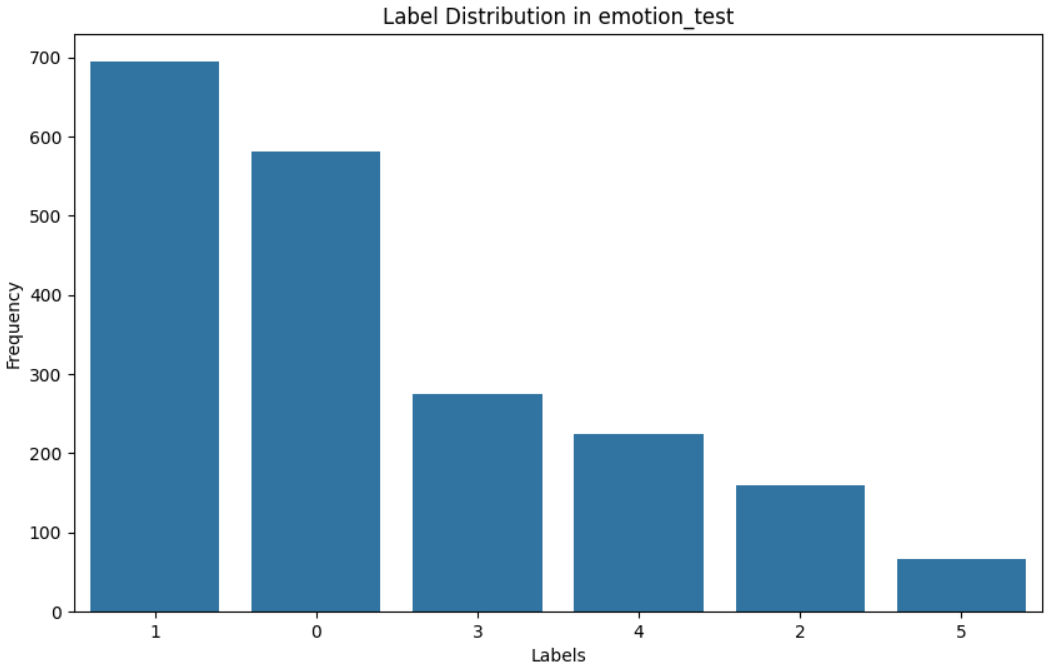
- Average: 19.15
- Standard deviation: 11.01
- Median: 17.0
- Max: 61
- Min: 3
- Quantiles (25%, 50%, 75%): 10.0, 17.0, 26.0

Stop Words Proportion

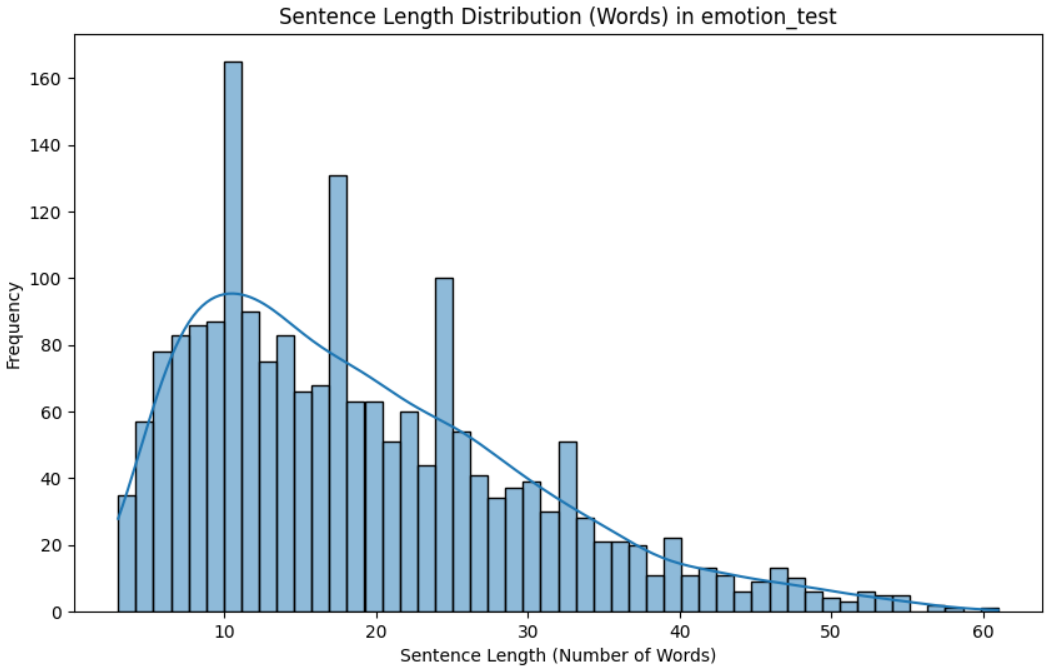
51.44%

Label Distribution

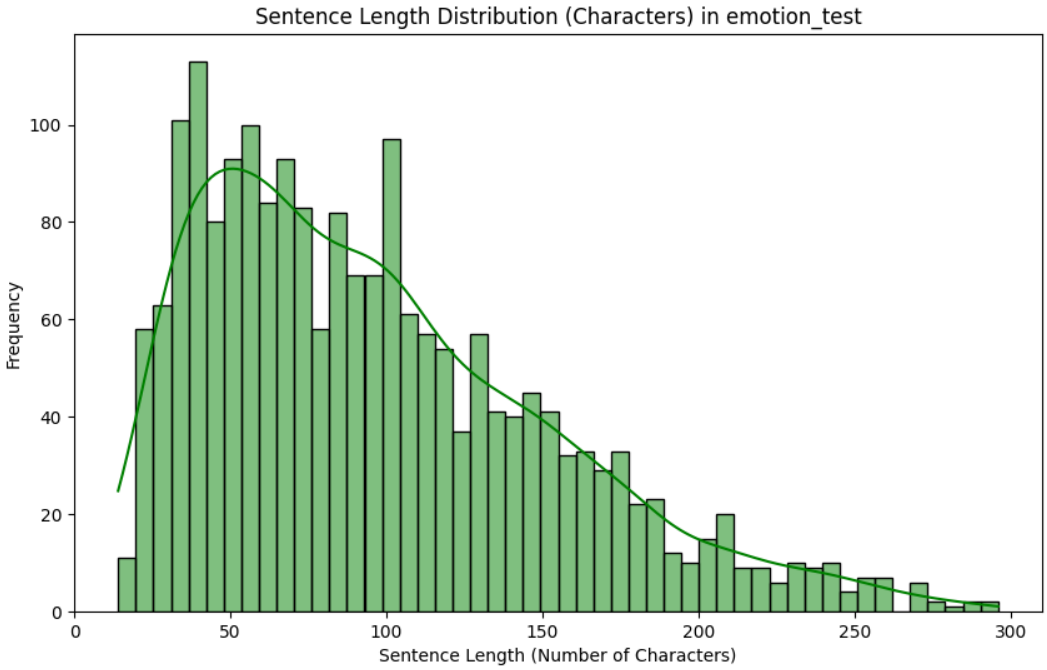
Label	Frequency
1	695
0	581
3	275
4	224
2	159
5	66



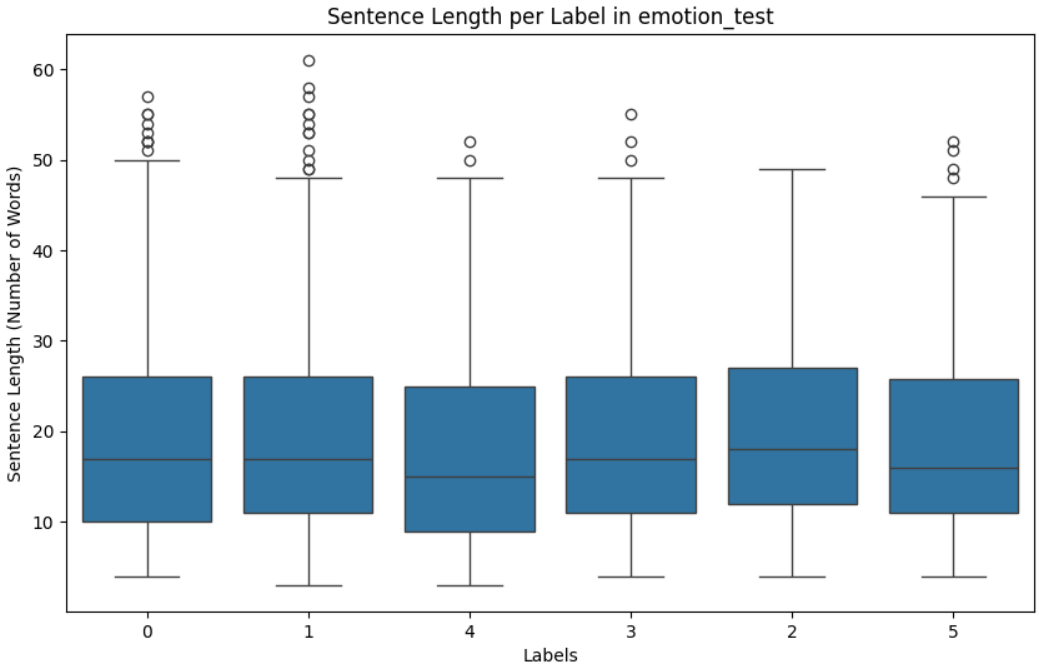
Sentence Length Distribution (Words)



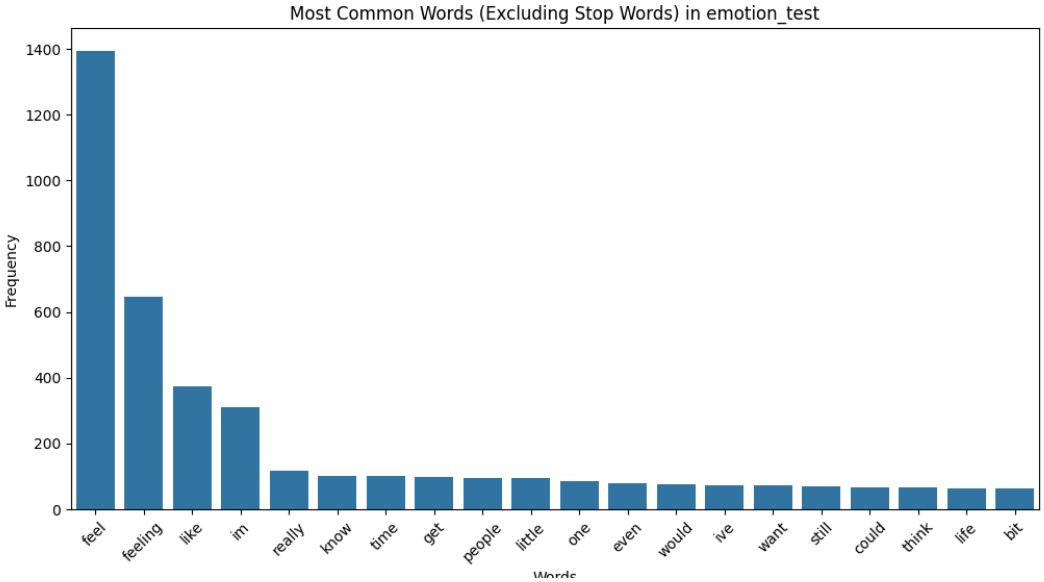
Sentence Length Distribution (Characters)



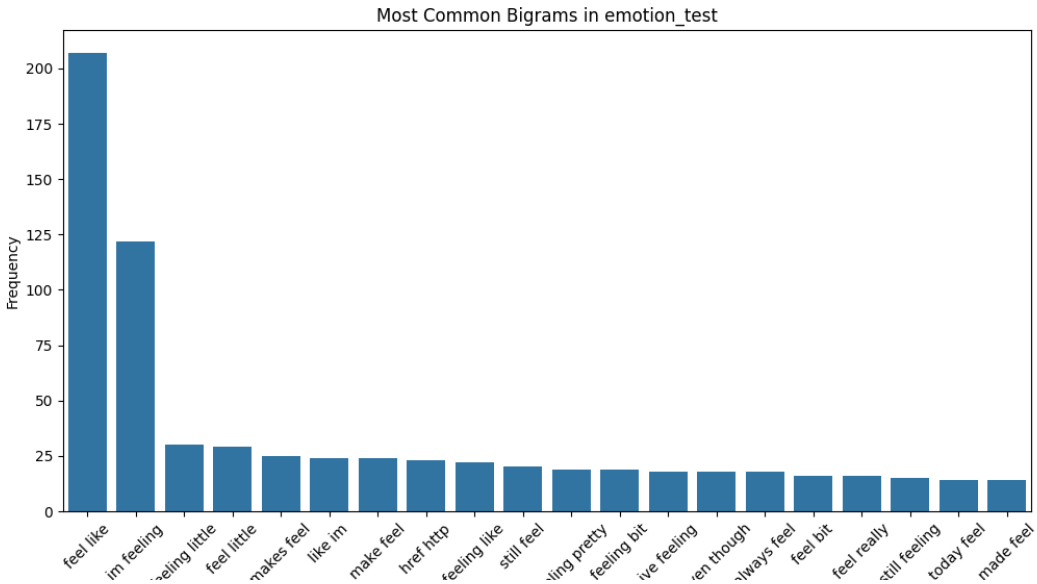
Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams



Analysis of emotion_train

Basic Information

- Number of samples: 16000
- Missing 'text' entries: 0
- Missing 'label' entries: 0

- Number of unique sentences: 15969
- Number of unique labels: 6
- Vocabulary size: 15212

Sentence Length (Words)

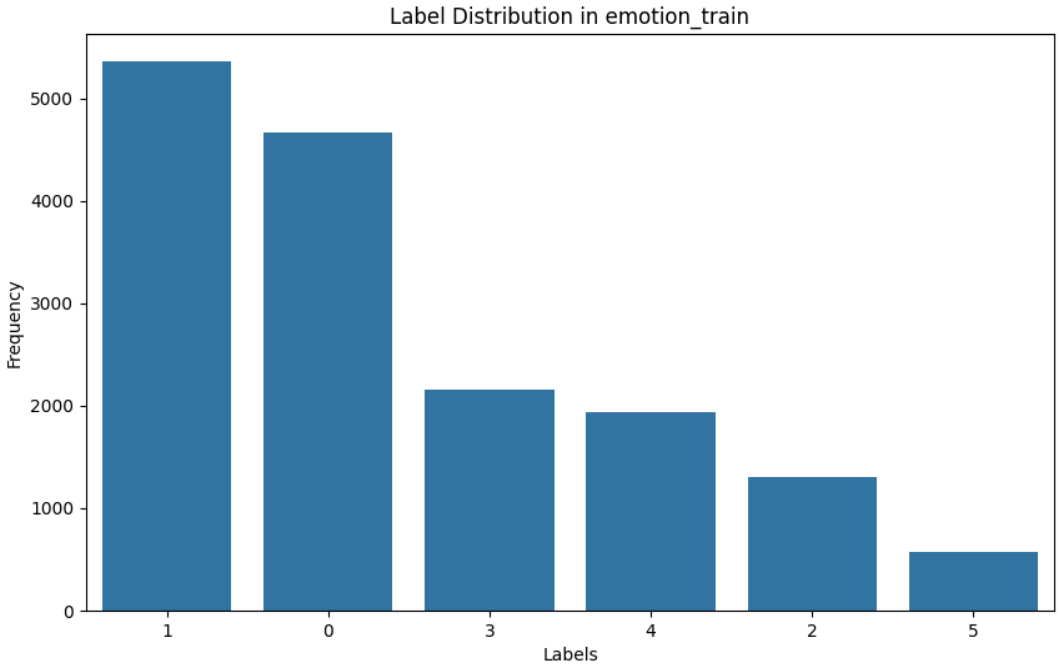
- Average: 19.17
- Standard deviation: 10.99
- Median: 17.0
- Max: 66
- Min: 2
- Quantiles (25%, 50%, 75%): 11.0, 17.0, 25.0

Stop Words Proportion

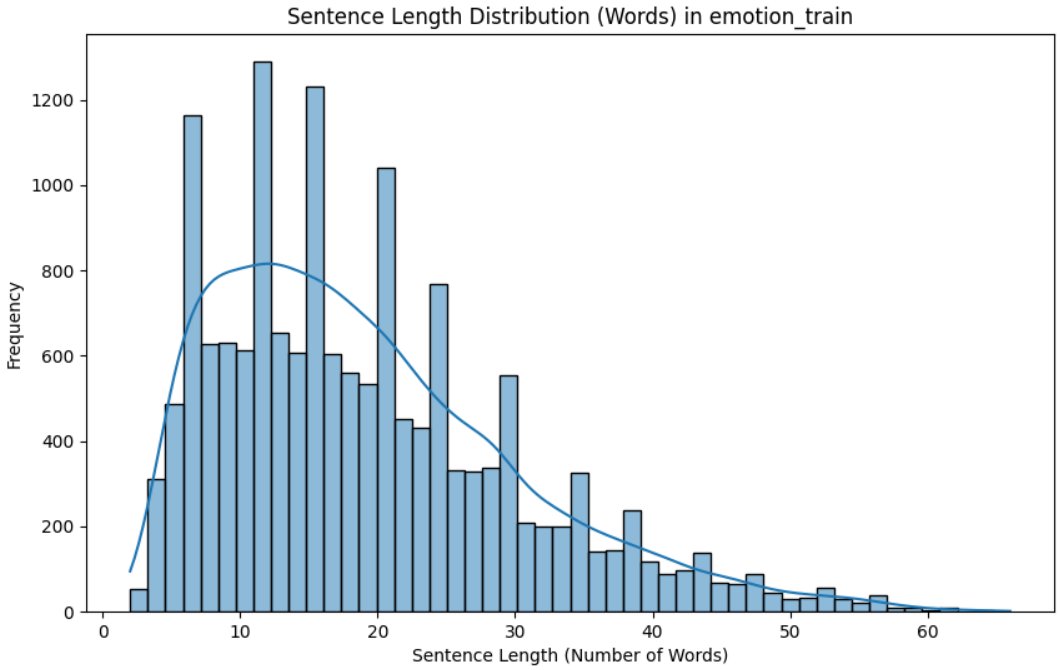
51.2%

Label Distribution

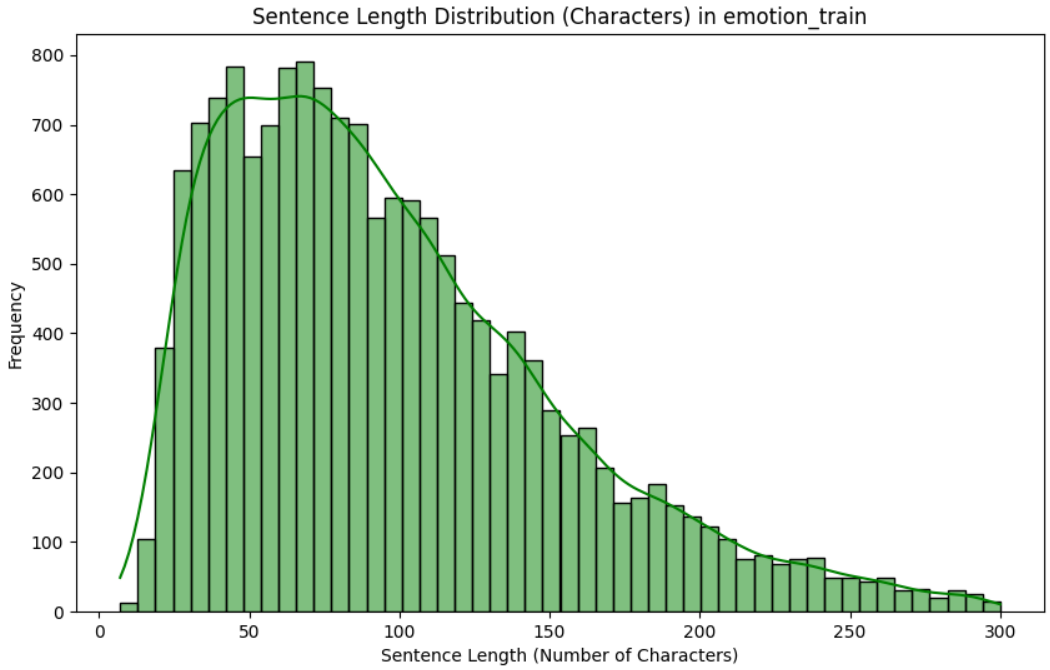
Label	Frequency
1	5362
0	4666
3	2159
4	1937
2	1304
5	572



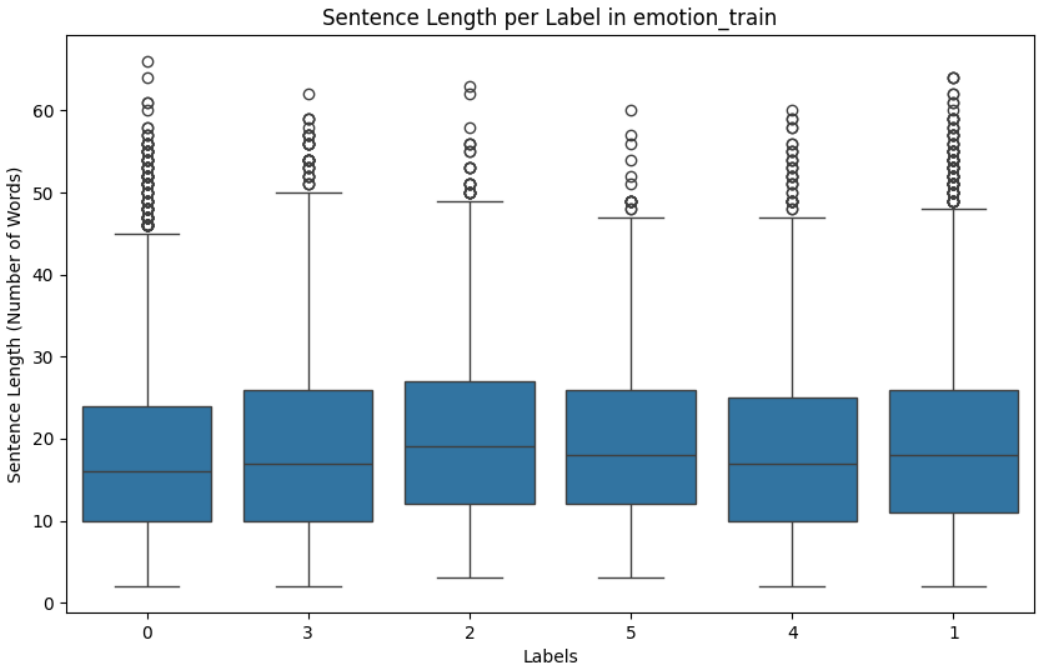
Sentence Length Distribution (Words)



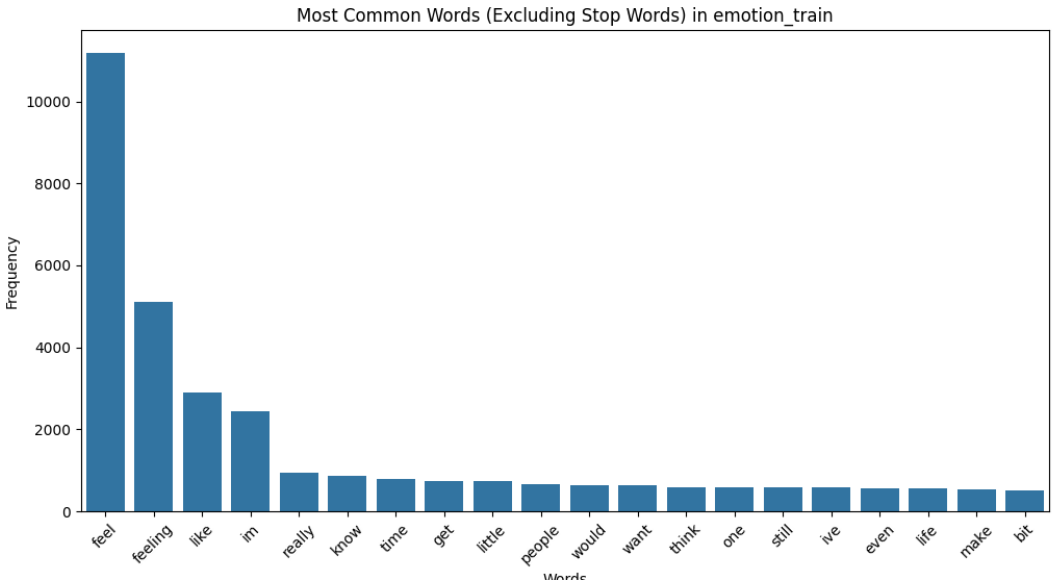
Sentence Length Distribution (Characters)



Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams