Dataset Dimensional Analysis Report

Generated by Python Script

October 21, 2024

Analysis of sst2_test

Basic Information

- Number of samples: 1821
- Missing 'text' entries: 0
- Missing 'label' entries: 0
- Number of unique sentences: 1821
- Number of unique labels: 2
- Vocabulary size: 7055

Sentence Length (Words)

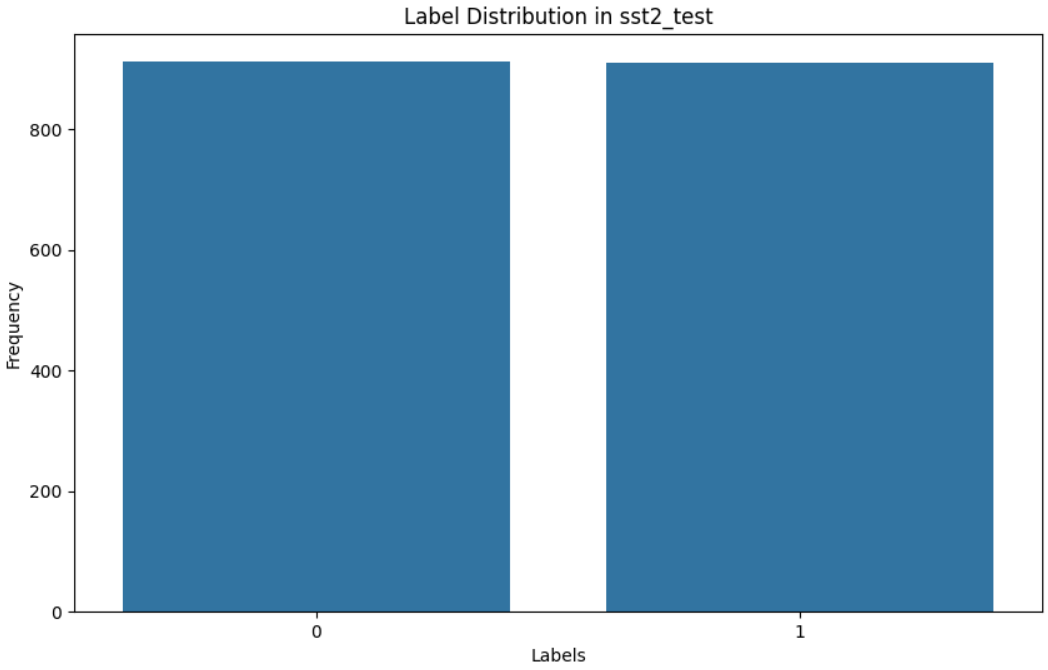
- Average: 19.23
- Standard deviation: 8.92
- Median: 18.0
- Max: 56
- Min: 2
- Quantiles (25%, 50%, 75%): 12.0, 18.0, 25.0

Stop Words Proportion

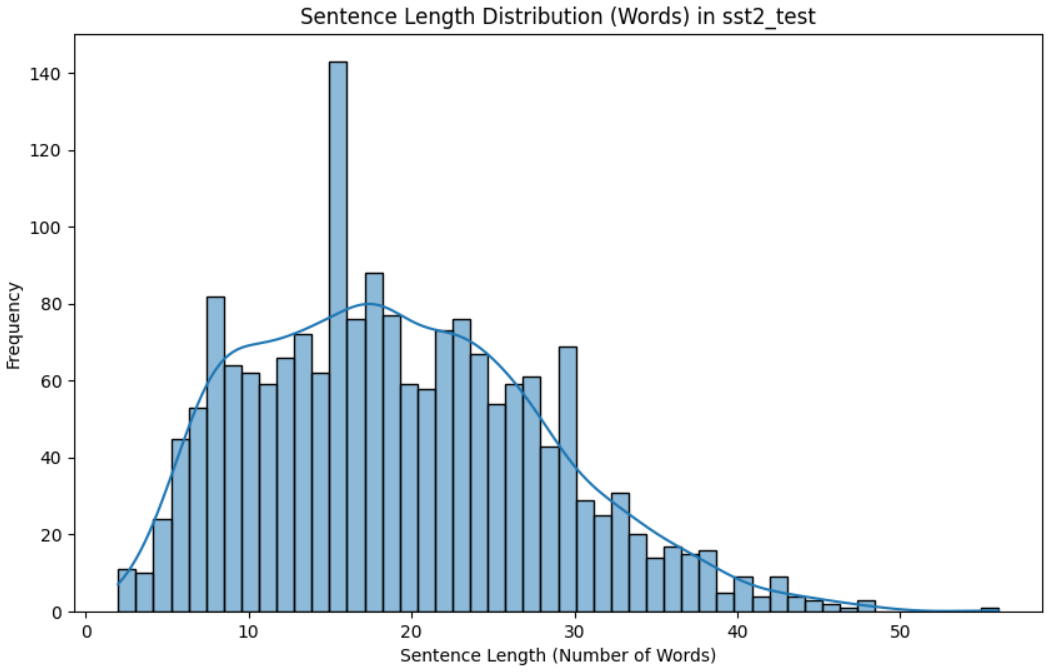
37.77%

Label Distribution

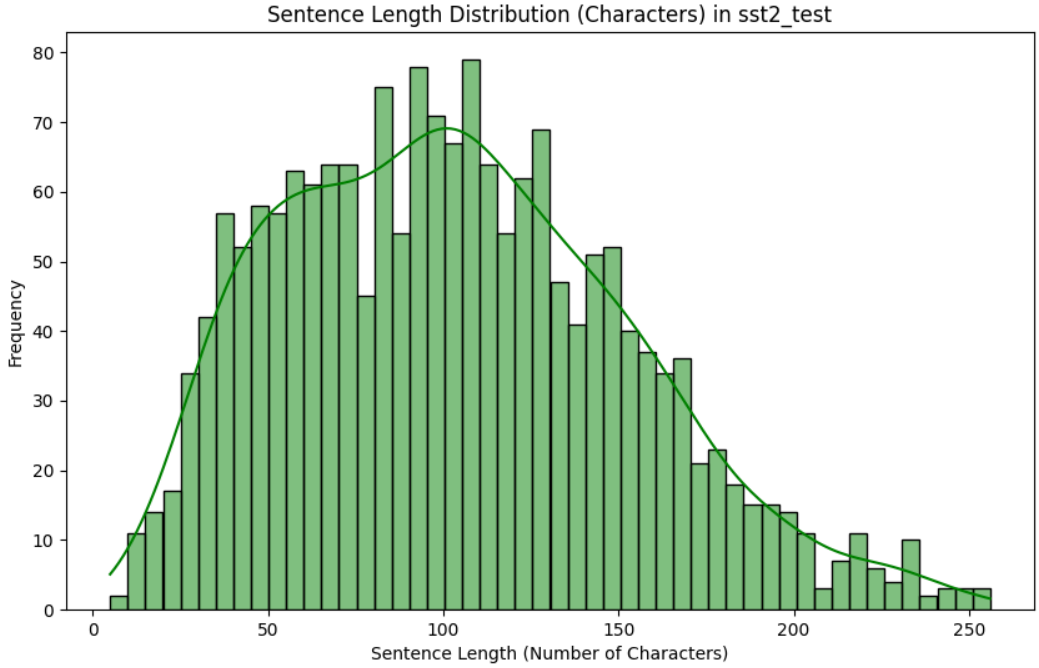
Label	Frequency
0	912
1	909



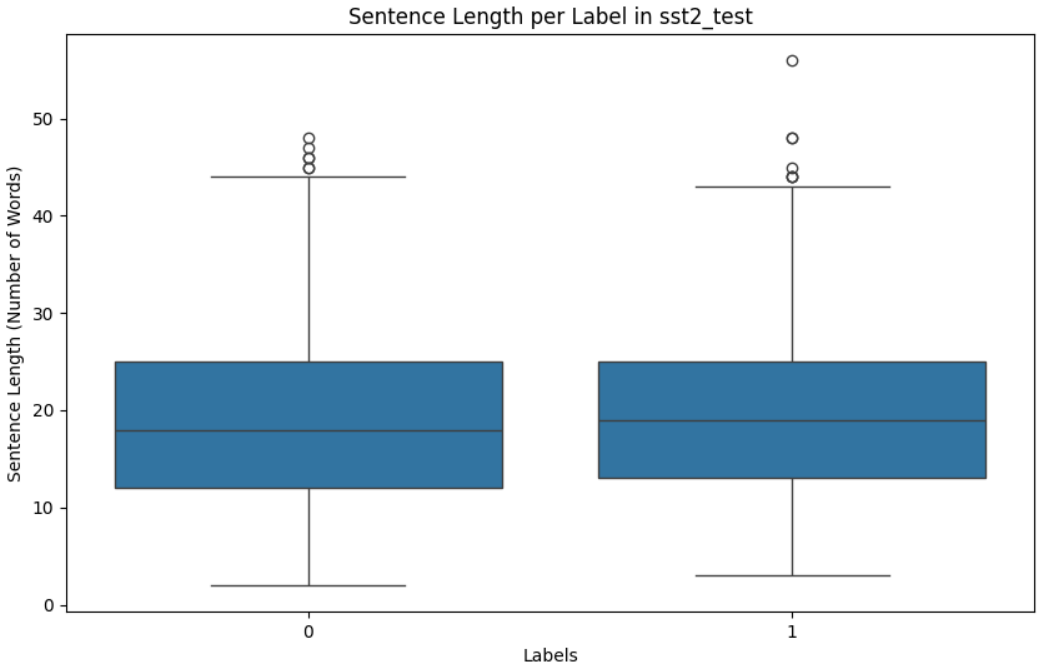
Sentence Length Distribution (Words)



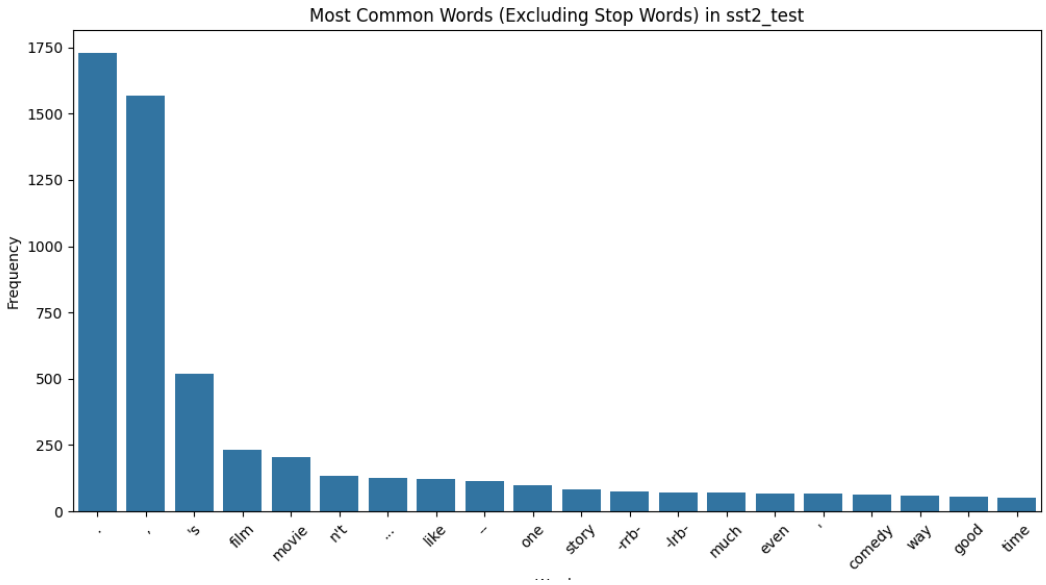
Sentence Length Distribution (Characters)



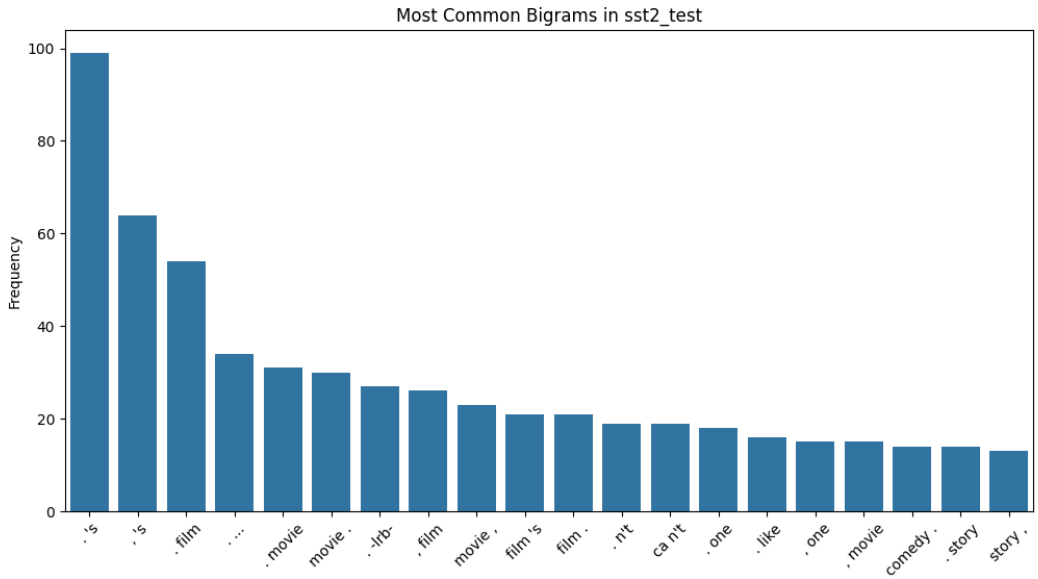
Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams



Analysis of sst2_train

Basic Information

- Number of samples: 6920
- Missing 'text' entries: 0
- Missing 'label' entries: 0

- Number of unique sentences: 6911
- Number of unique labels: 2
- Vocabulary size: 14828

Sentence Length (Words)

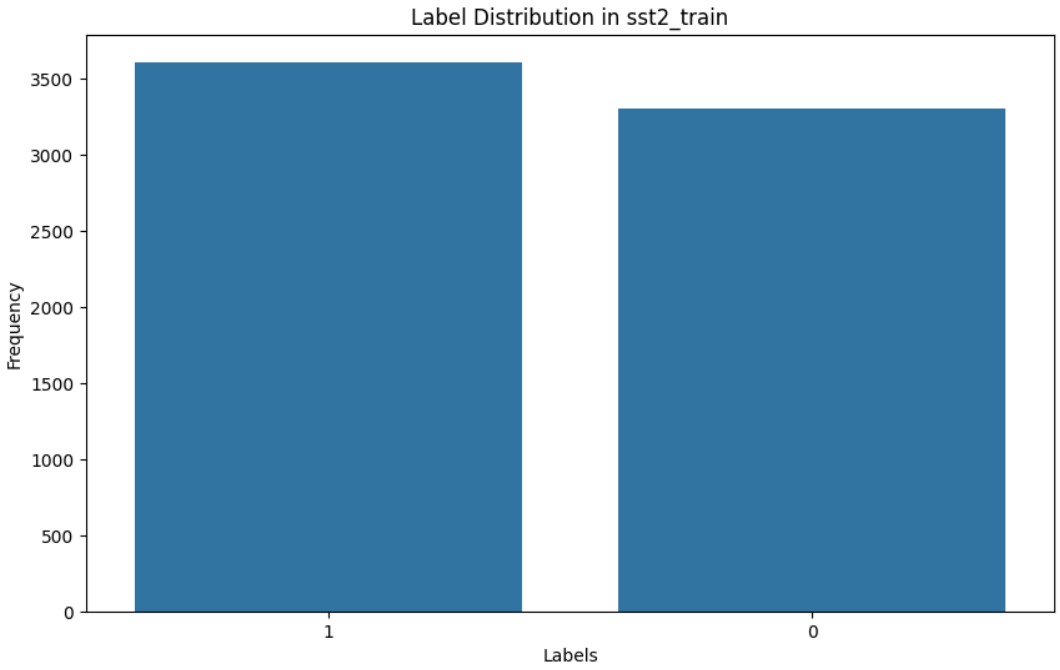
- Average: 19.3
- Standard deviation: 9.32
- Median: 19.0
- Max: 52
- Min: 2
- Quantiles (25%, 50%, 75%): 12.0, 19.0, 25.0

Stop Words Proportion

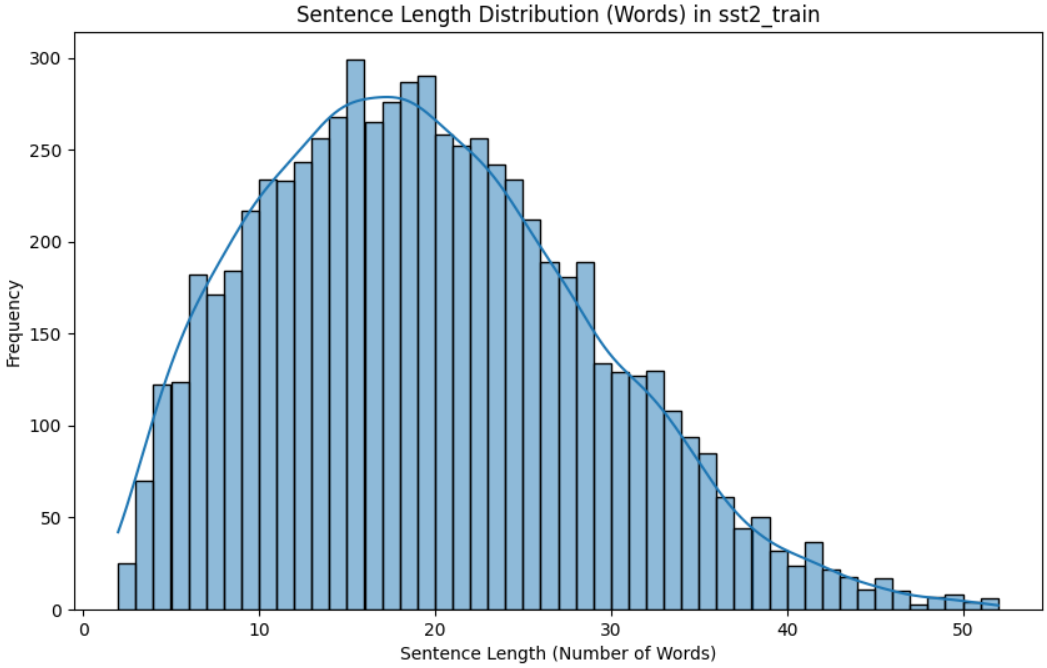
37.48%

Label Distribution

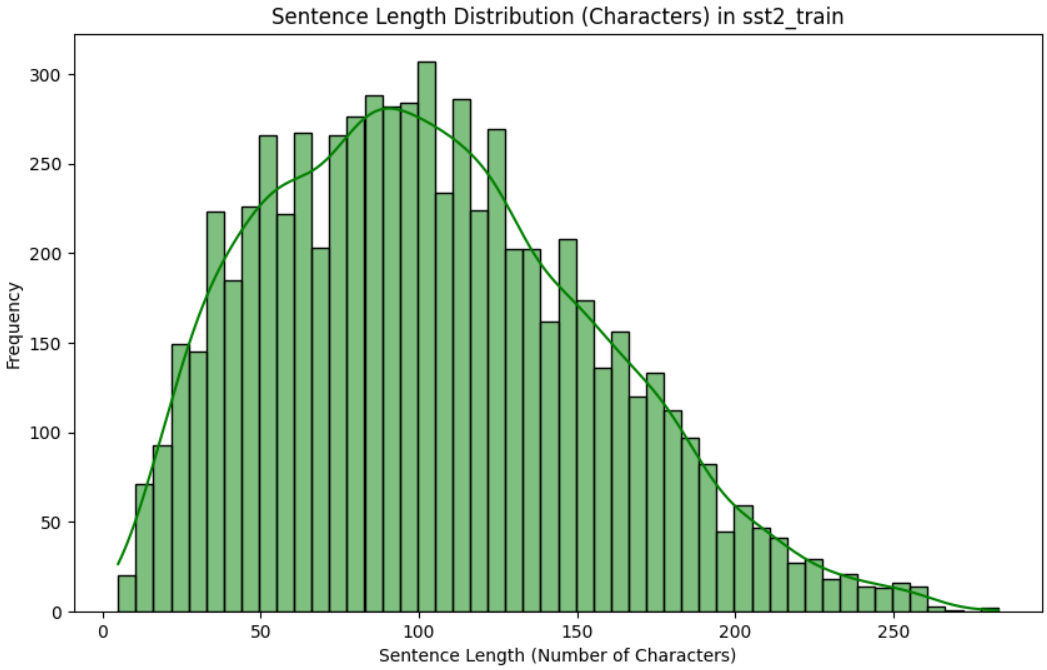
Label	Frequency
1	3610
0	3310



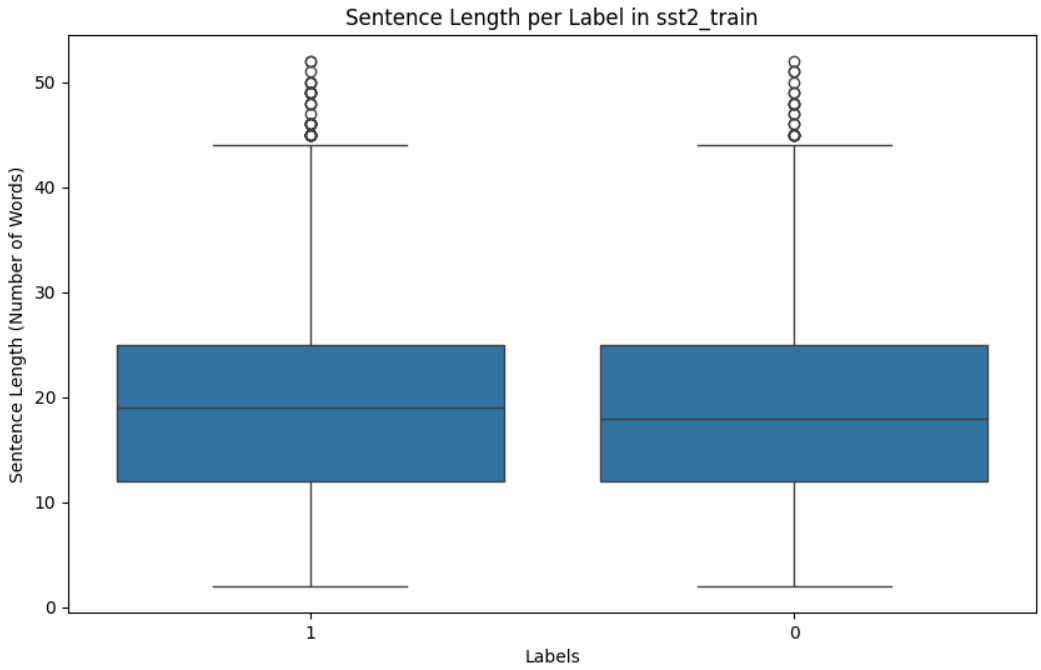
Sentence Length Distribution (Words)



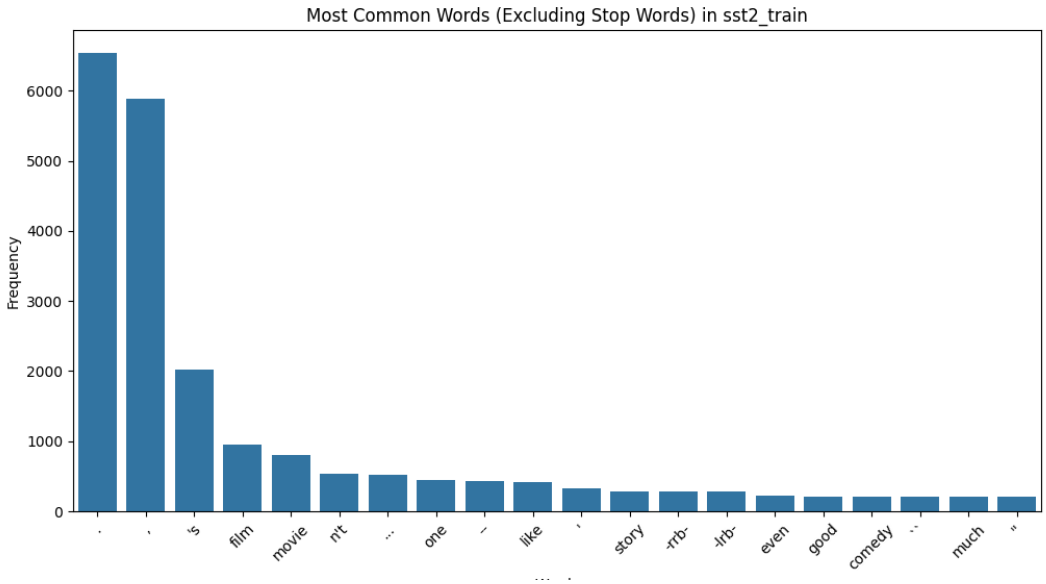
Sentence Length Distribution (Characters)



Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams