Dataset Dimensional Analysis Report

Generated by Python Script

October 21, 2024

Analysis of imdb_test_reduce

Basic Information

- Number of samples: 7000
- Missing 'text' entries: 0
- Missing 'label' entries: 0
- Number of unique sentences: 6937
- Number of unique labels: 2
- Vocabulary size: 66830

Sentence Length (Words)

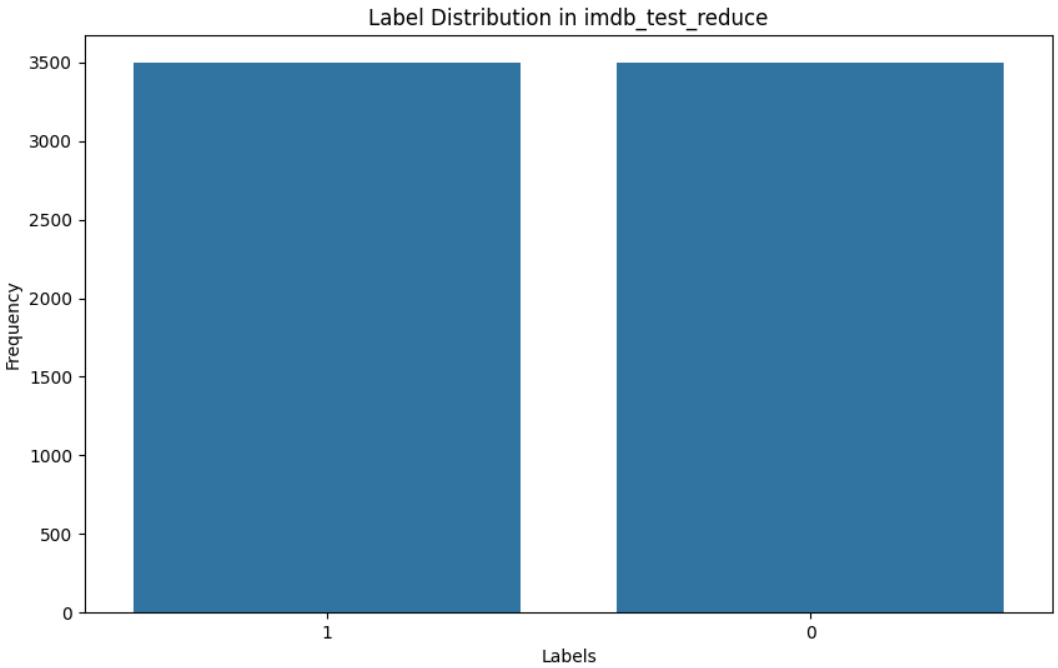
- Average: 95.56
- Standard deviation: 28.46
- Median: 105.0
- Max: 131
- Min: 4
- Quantiles (25%, 50%, 75%): 72.0, 105.0, 120.0

Stop Words Proportion

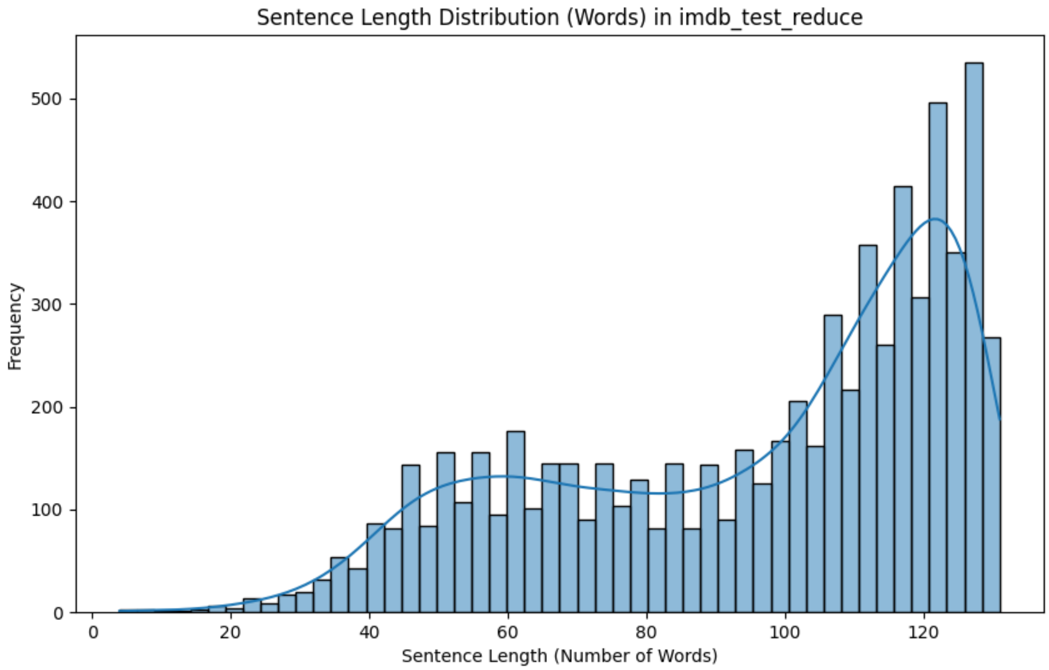
44.86%

Label Distribution

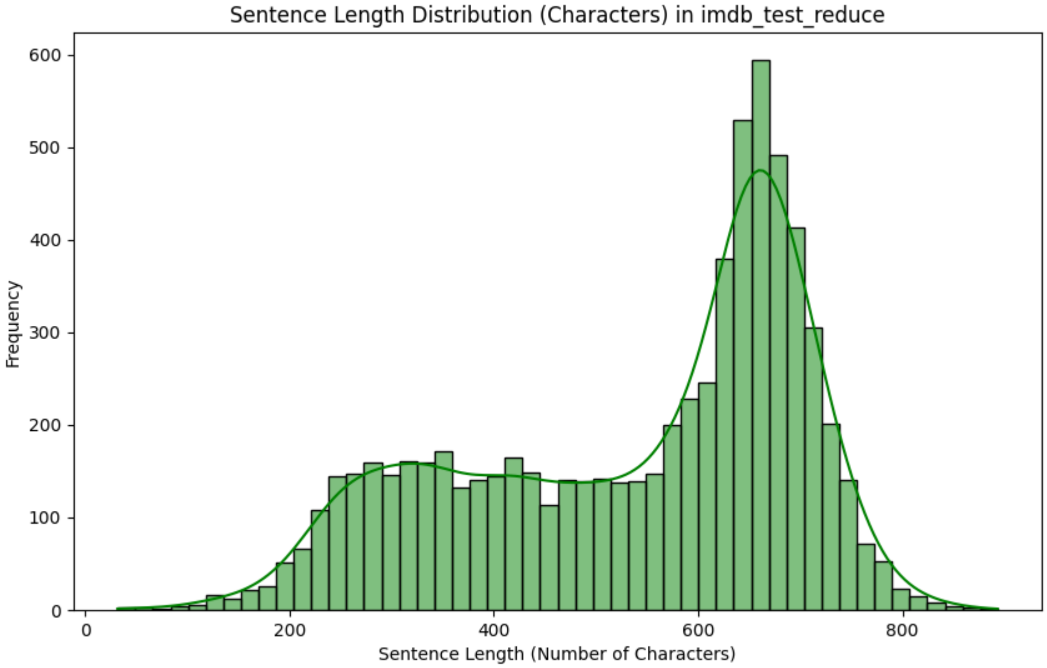
Label	Frequency
1	3500
0	3500



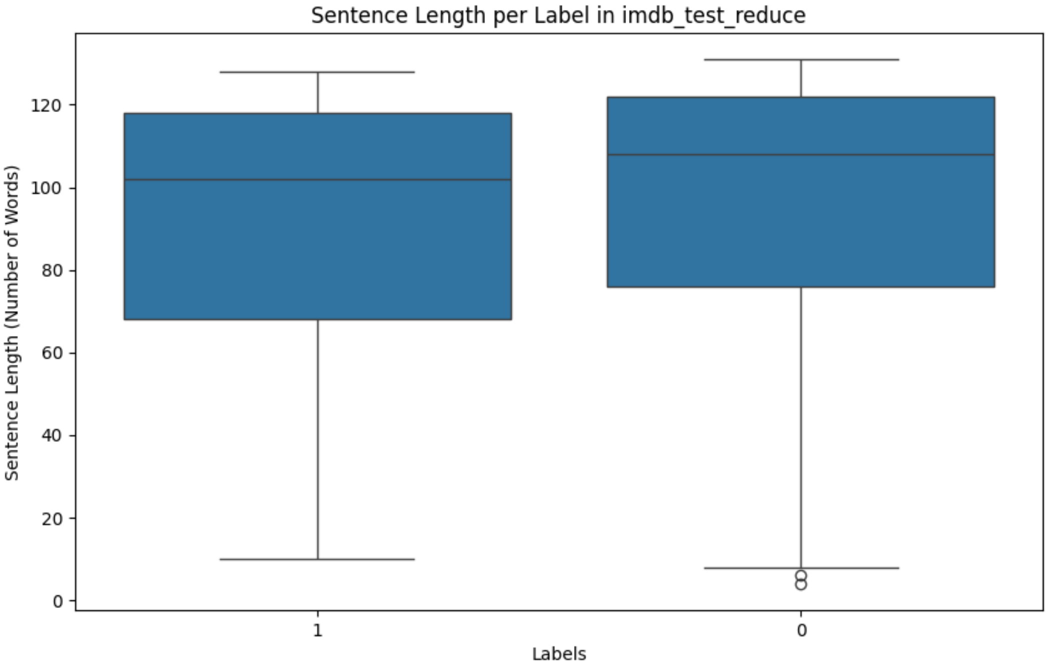
Sentence Length Distribution (Words)



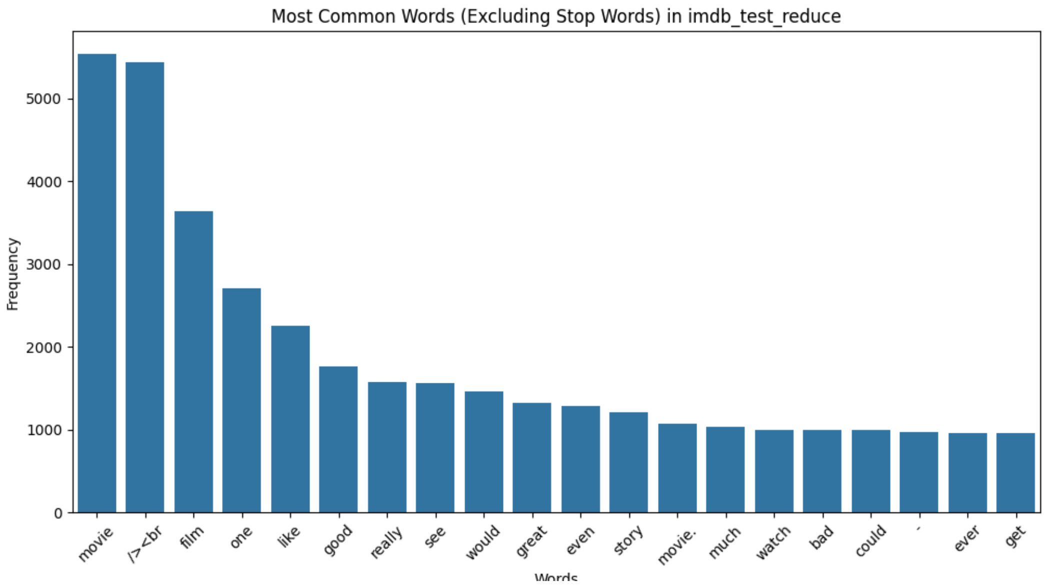
Sentence Length Distribution (Characters)



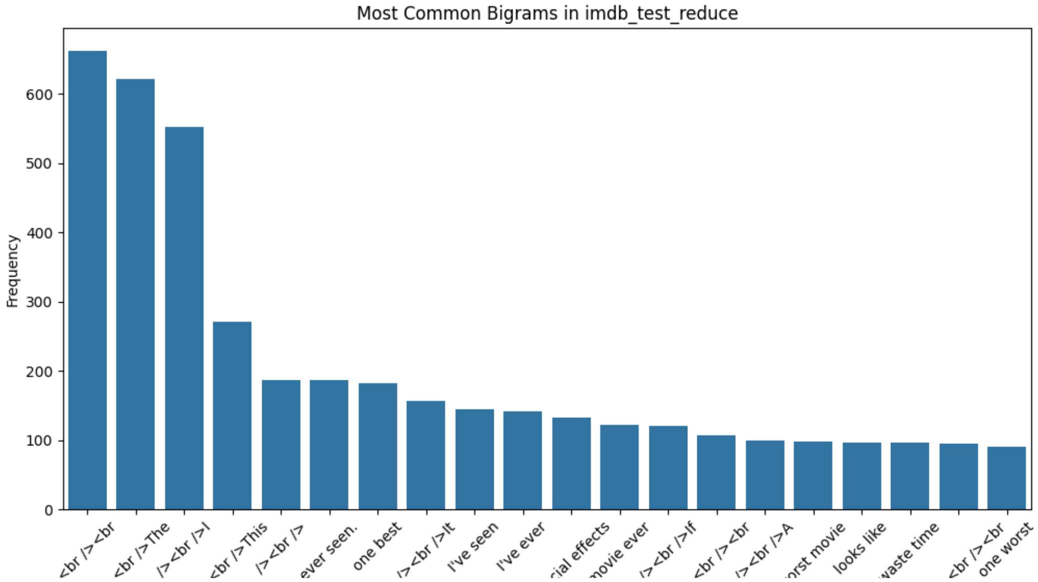
Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams



Analysis of imdb_train_reduce

Basic Information

- Number of samples: 6948
- Missing 'text' entries: 0
- Missing 'label' entries: 0

- Number of unique sentences: 6911
- Number of unique labels: 2
- Vocabulary size: 67343

Sentence Length (Words)

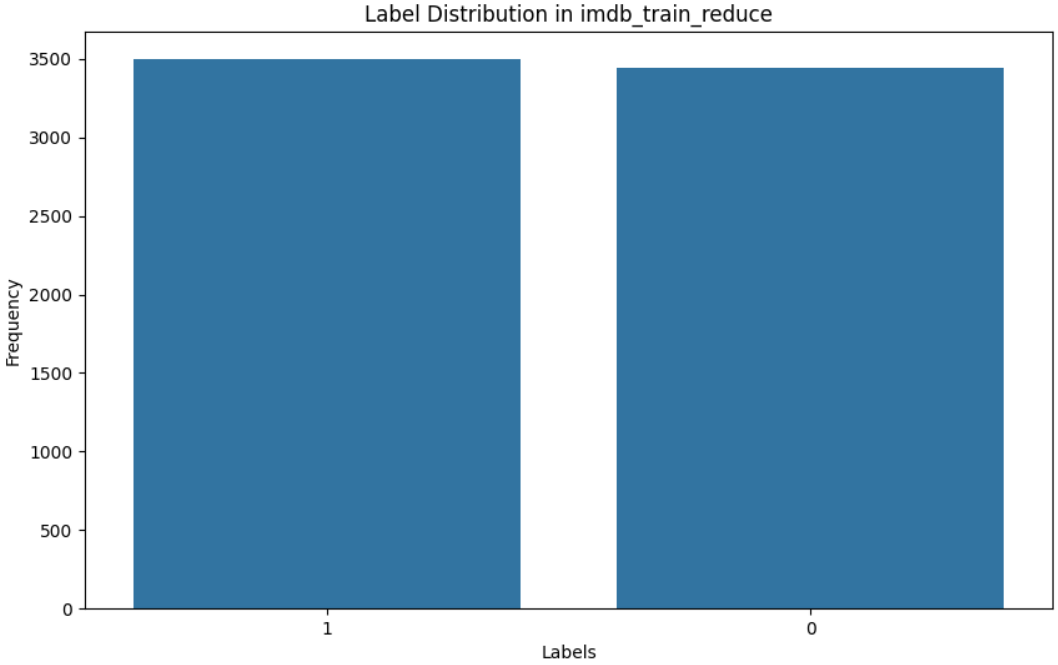
- Average: 97.12
- Standard deviation: 28.26
- Median: 108.0
- Max: 132
- Min: 10
- Quantiles (25%, 50%, 75%): 75.0, 108.0, 121.0

Stop Words Proportion

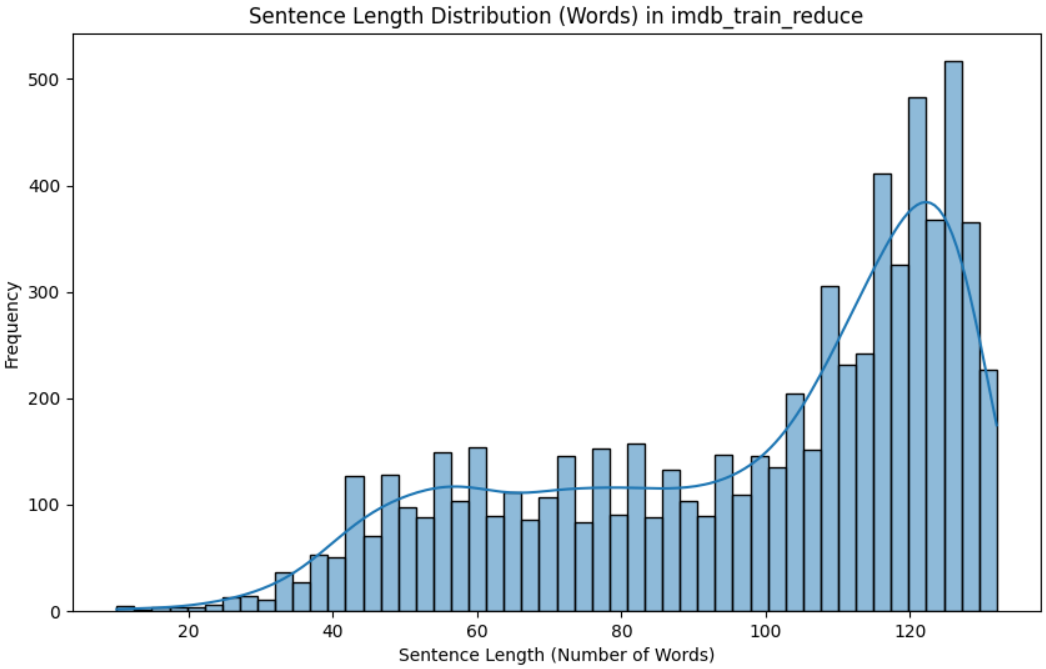
44.89%

Label Distribution

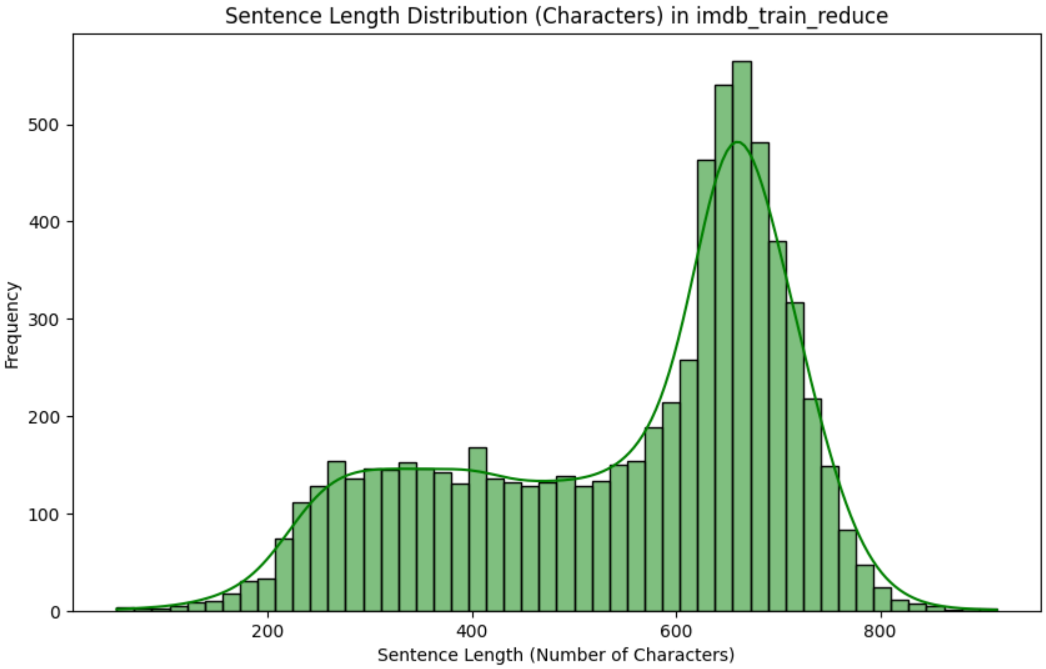
Label	Frequency
1	3500
0	3448



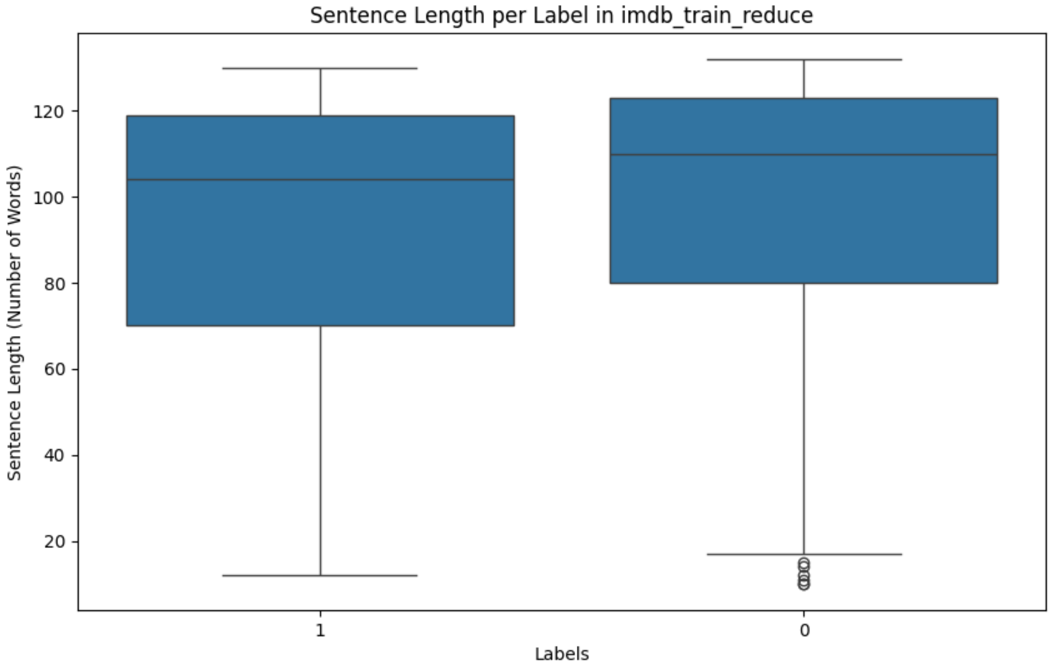
Sentence Length Distribution (Words)



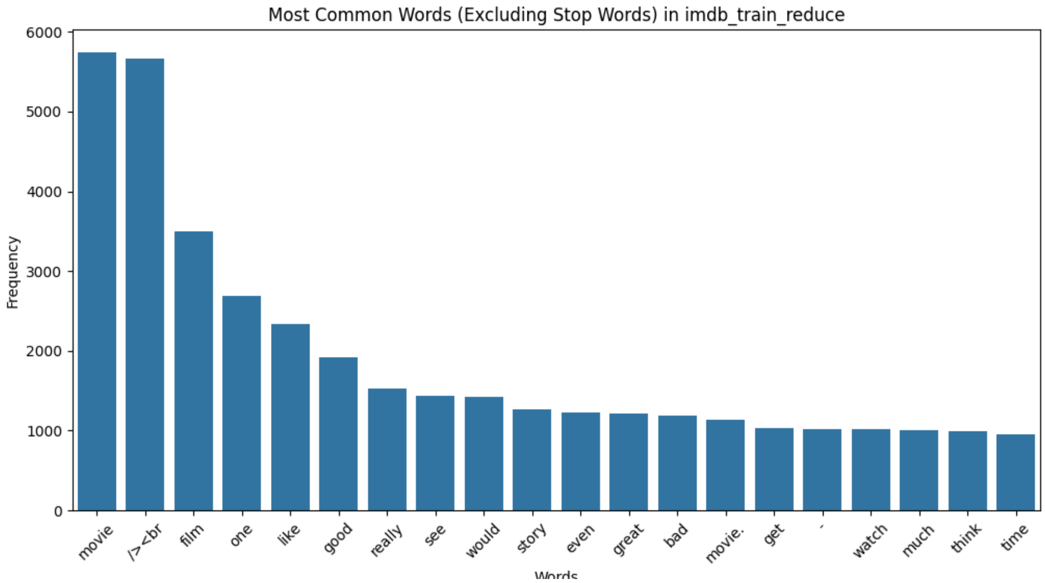
Sentence Length Distribution (Characters)



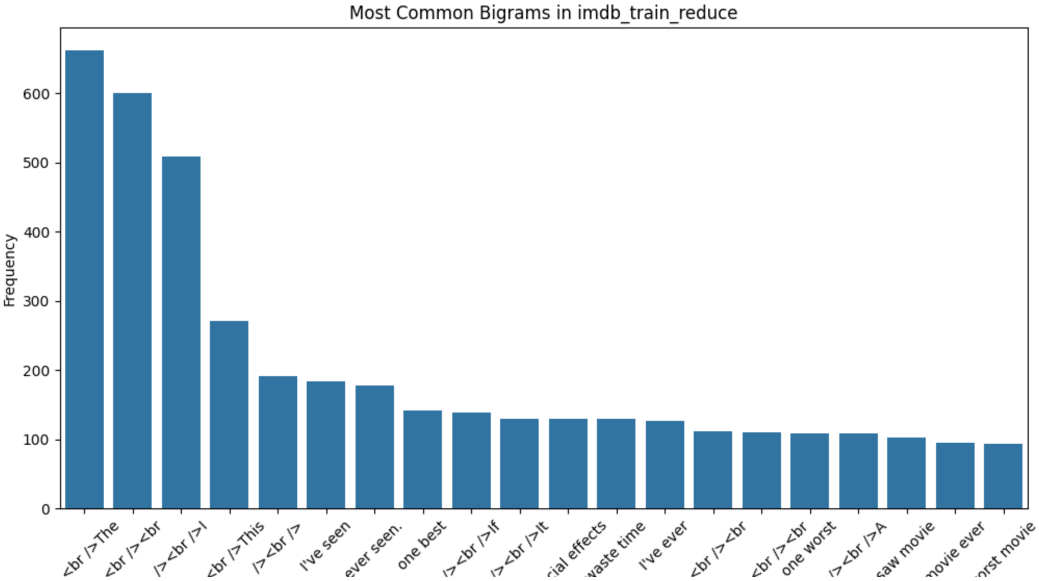
Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams


Most Common Bigrams in imdb_train_reduce
Frequency
<br />The
<br />
<br />I
<br />This
/><br />
I've seen
ever seen.
one best
/><br />If
/><br />It
cial effects
waste time
I've ever
<br /><br
<br /><br
one worst
/><br />A
saw movie
movie ever
orst movie