Dataset Dimensional Analysis Report

Generated by Python Script

October 21, 2024

Analysis of imdb_test

Basic Information

- Number of samples: 25000
- Missing 'text' entries: 0
- Missing 'label' entries: 0
- Number of unique sentences: 24801
- Number of unique labels: 2
- Vocabulary size: 276678

Sentence Length (Words)

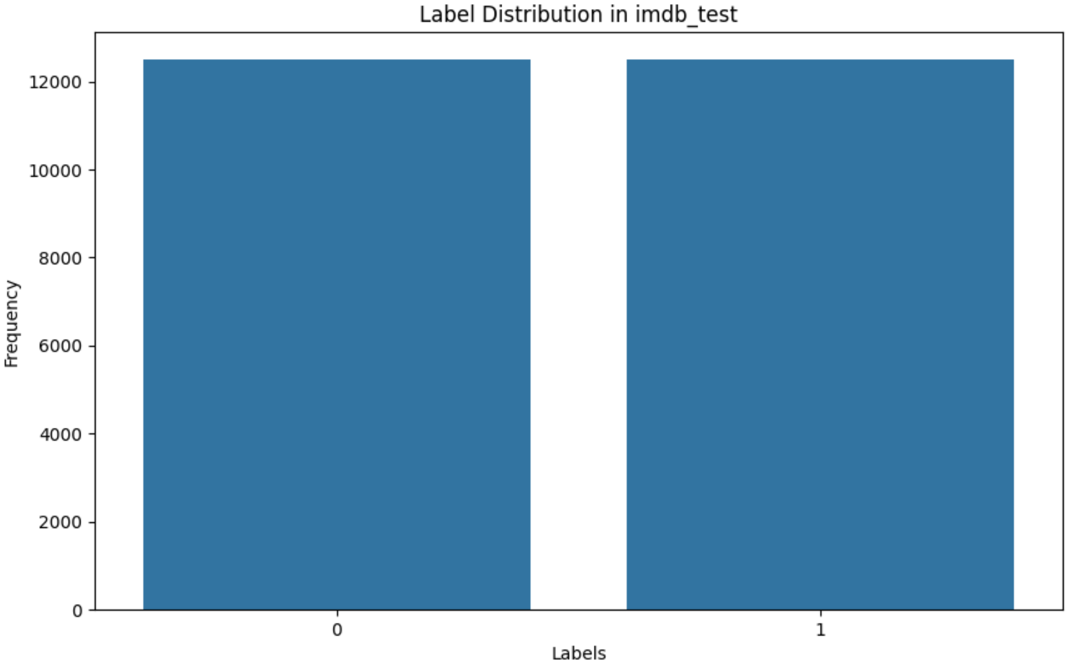
- Average: 228.53
- Standard deviation: 168.88
- Median: 172.0
- Max: 2278
- Min: 4
- Quantiles (25%, 50%, 75%): 126.0, 172.0, 277.0

Stop Words Proportion

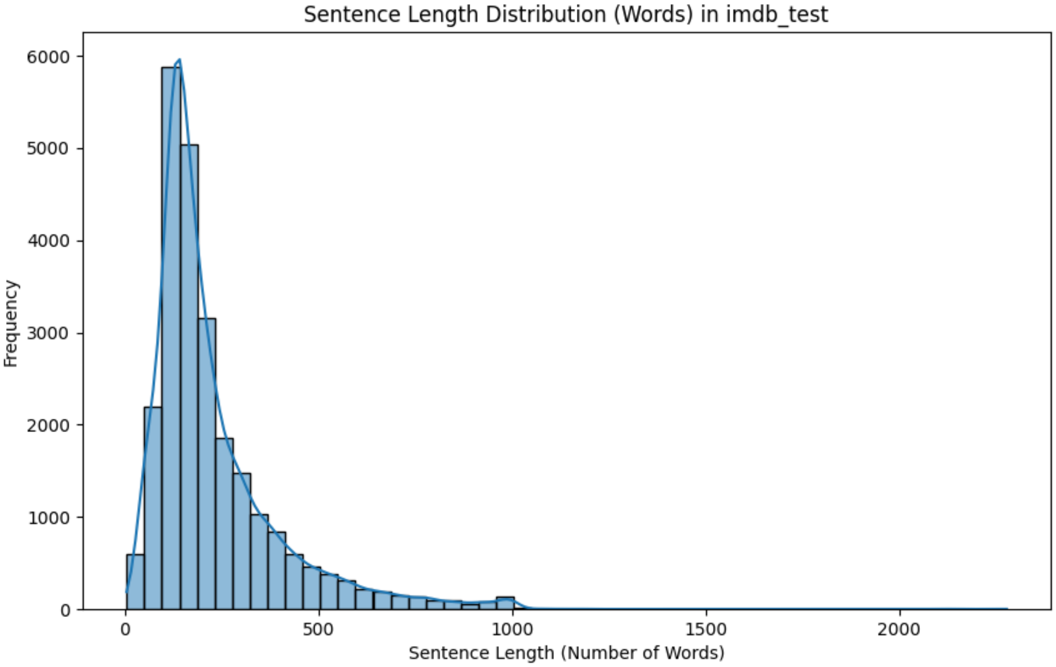
44.93%

Label Distribution

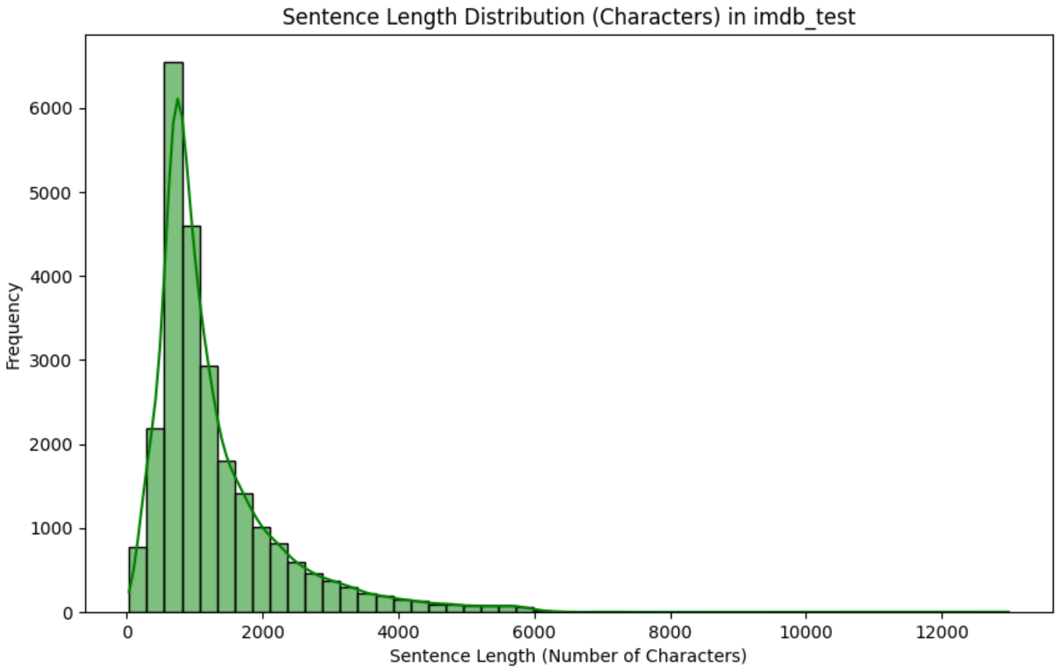
Label	Frequency
0	12500
1	12500



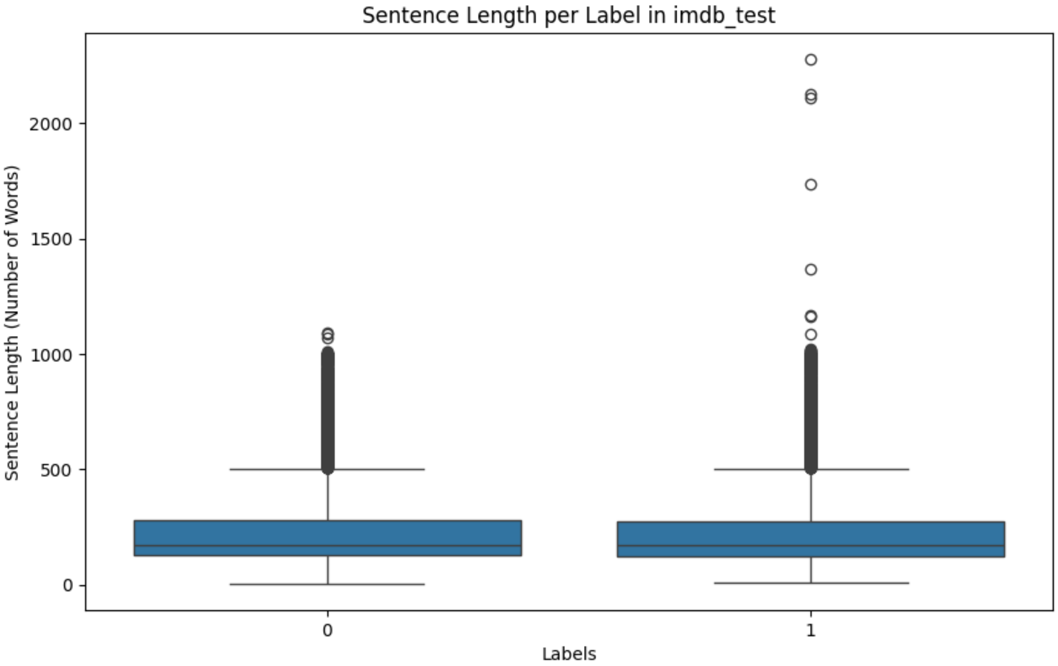
Sentence Length Distribution (Words)



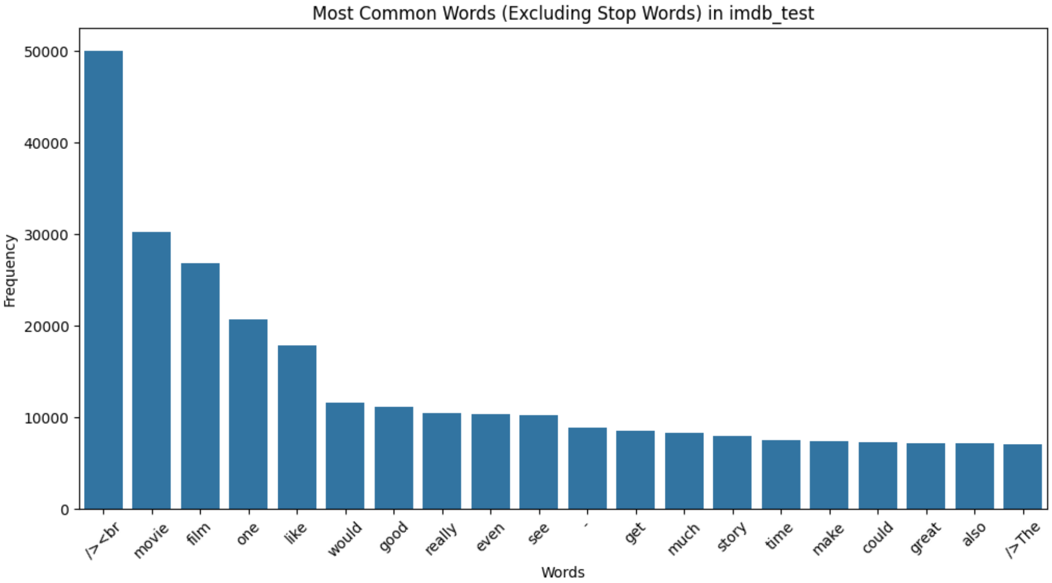
Sentence Length Distribution (Characters)



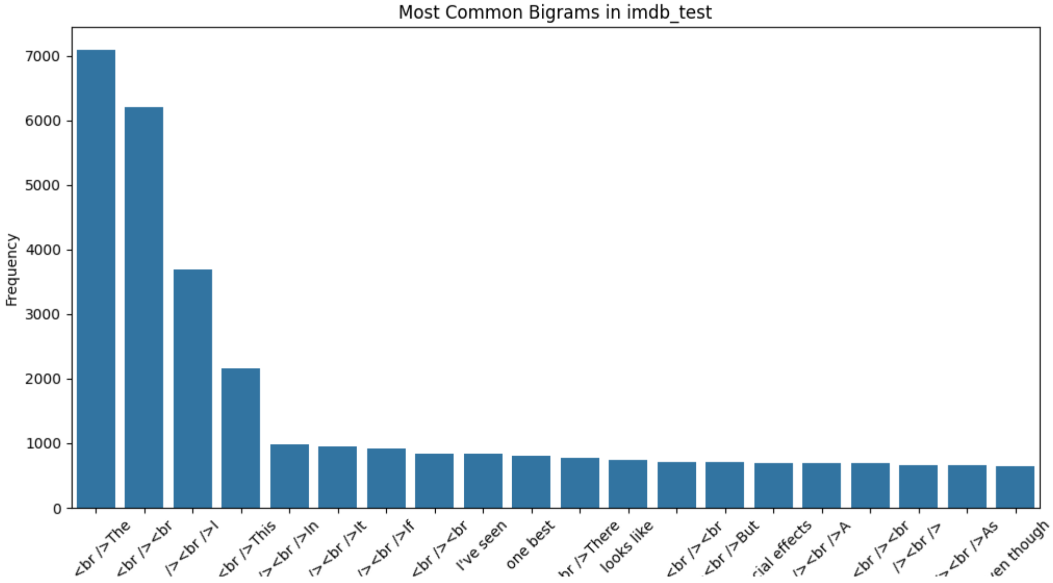
Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams



Analysis of imdb_train

Basic Information

- Number of samples: 25000
- Missing 'text' entries: 0
- Missing 'label' entries: 0

- Number of unique sentences: 24904
- Number of unique labels: 2
- Vocabulary size: 280617

Sentence Length (Words)

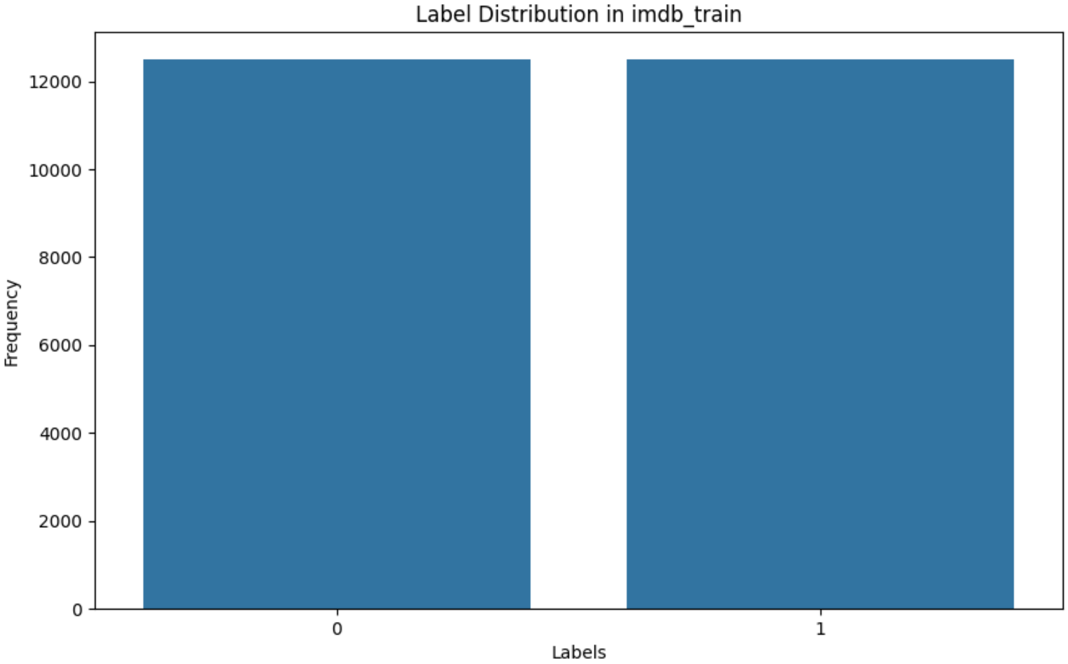
- Average: 233.79
- Standard deviation: 173.73
- Median: 174.0
- Max: 2470
- Min: 10
- Quantiles (25%, 50%, 75%): 127.0, 174.0, 284.0

Stop Words Proportion

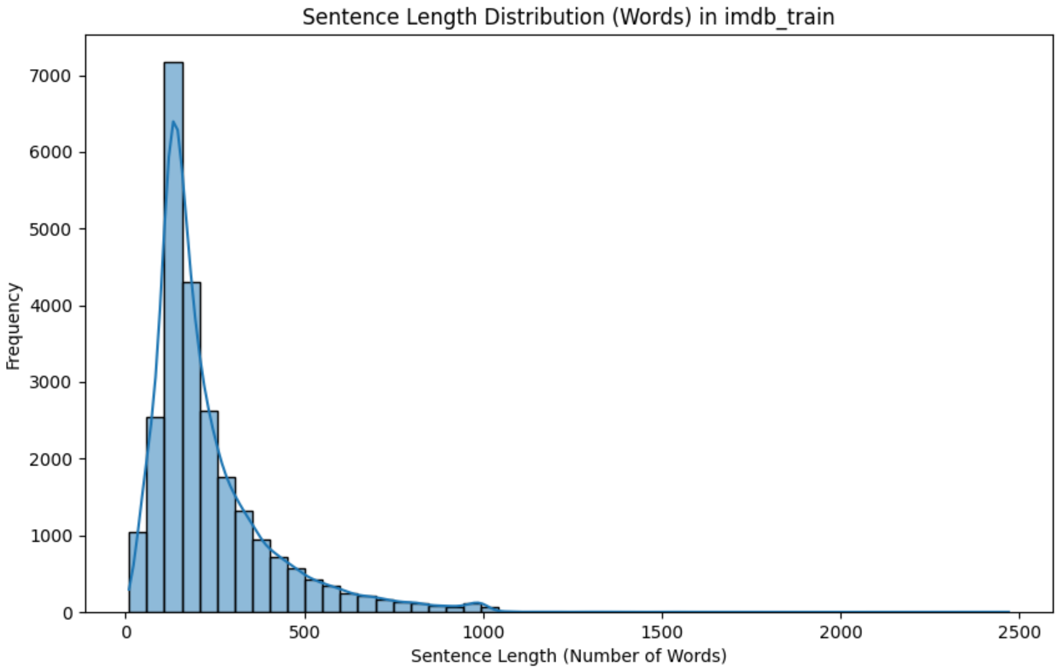
44.92%

Label Distribution

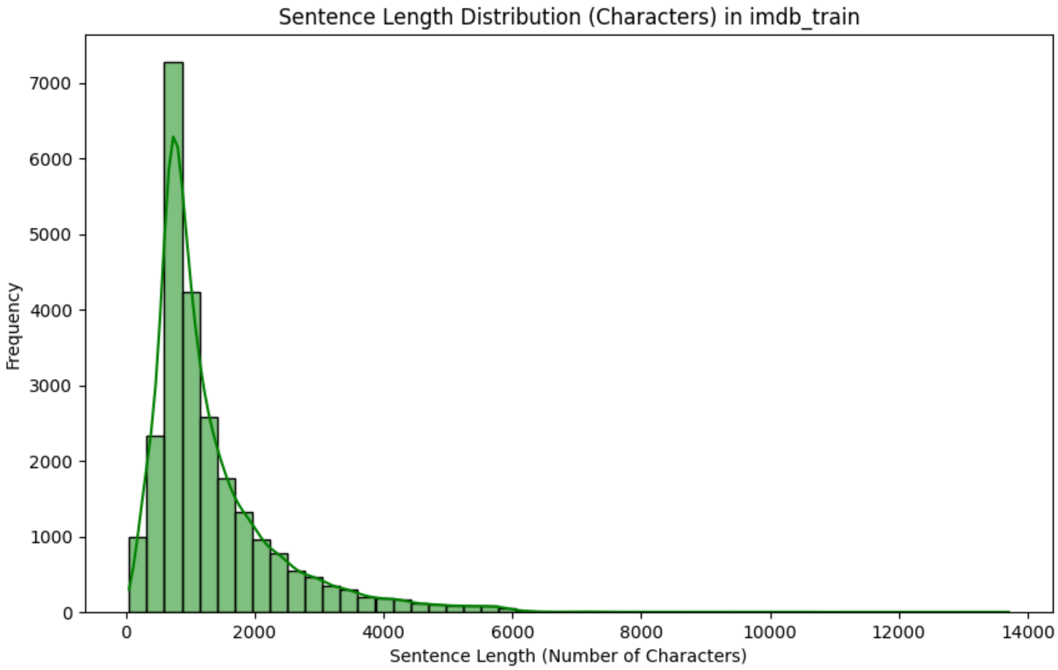
Label	Frequency
0	12500
1	12500



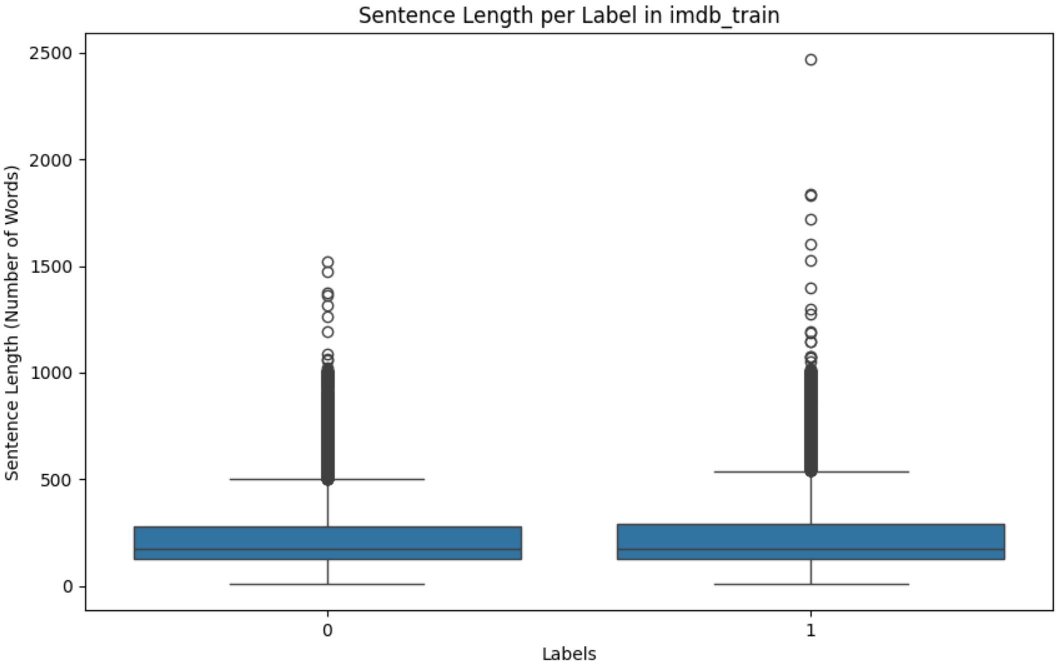
Sentence Length Distribution (Words)



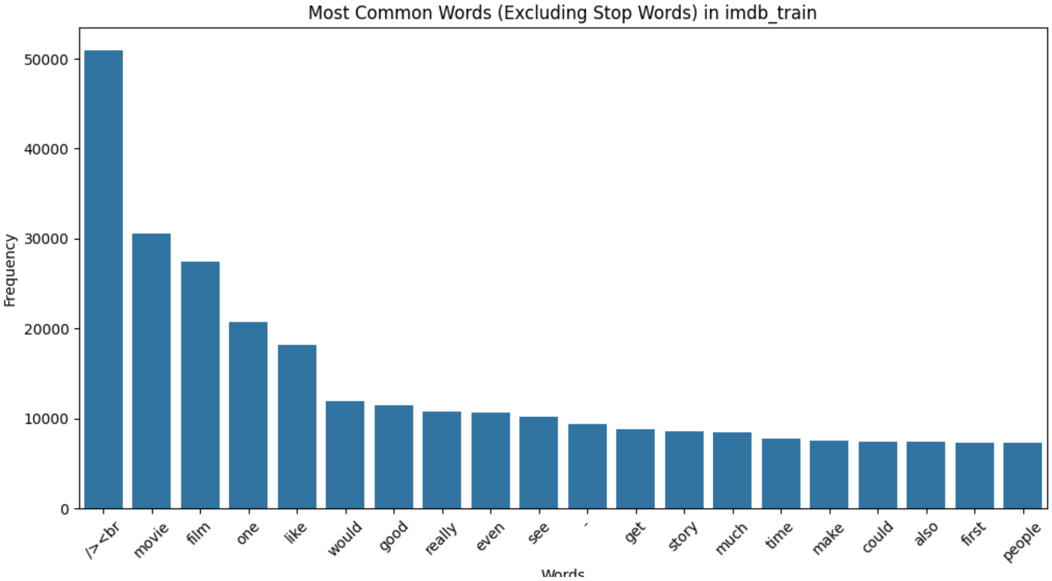
Sentence Length Distribution (Characters)



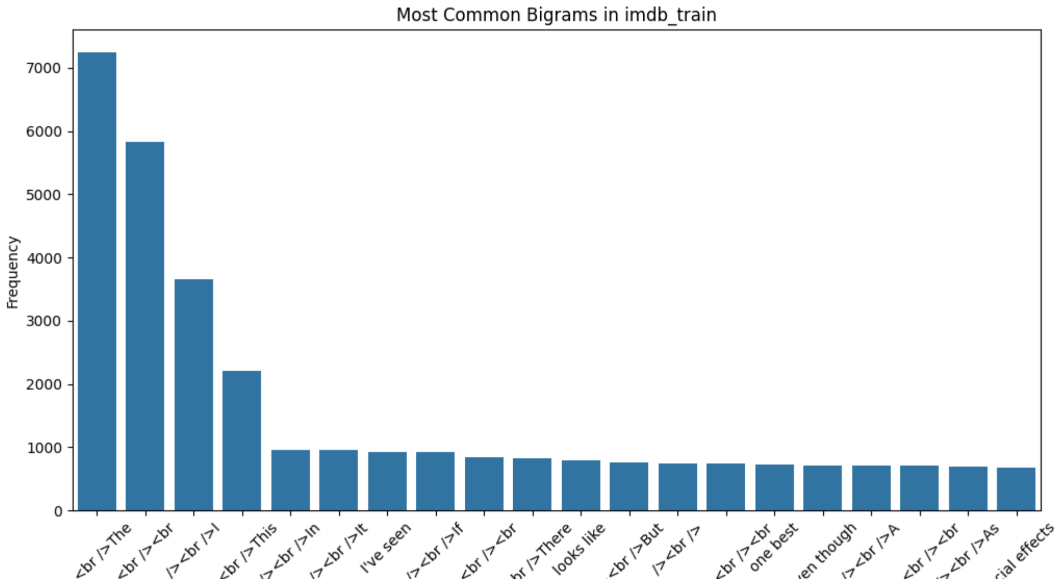
Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams


Most Common Bigrams in imdb_train
Frequency
<br />The
<br />
<br />I
<br />This
/><br />In
/><br />It
I've seen
/><br />If
<br /><br
br />There
looks like
<br />But
/><br />
<br /><br
one best
en though
/><br />A
<br />
<br />As
cial effects