Dataset Dimensional Analysis Report

Generated by Python Script

October 21, 2024

Analysis of rotten_tomatoes_test

Basic Information

- Number of samples: 1066
- Missing 'text' entries: 0
- Missing 'label' entries: 0
- Number of unique sentences: 1066
- Number of unique labels: 2
- Vocabulary size: 5578

Sentence Length (Words)

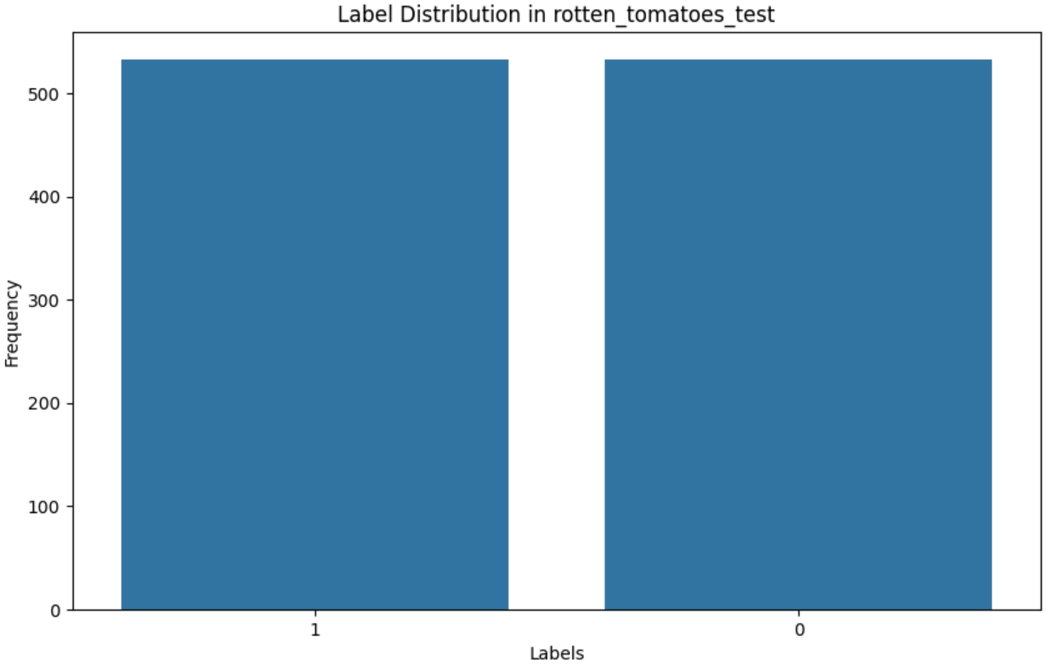
- Average: 21.22
- Standard deviation: 9.52
- Median: 20.0
- Max: 52
- Min: 3
- Quantiles (25%, 50%, 75%): 14.0, 20.0, 27.75

Stop Words Proportion

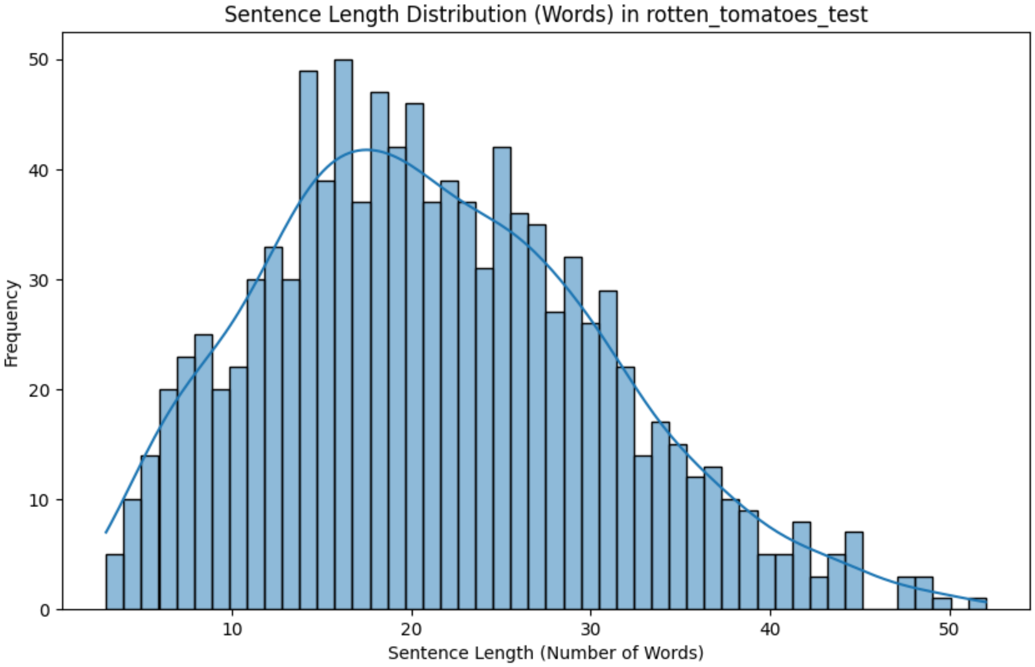
37.65%

Label Distribution

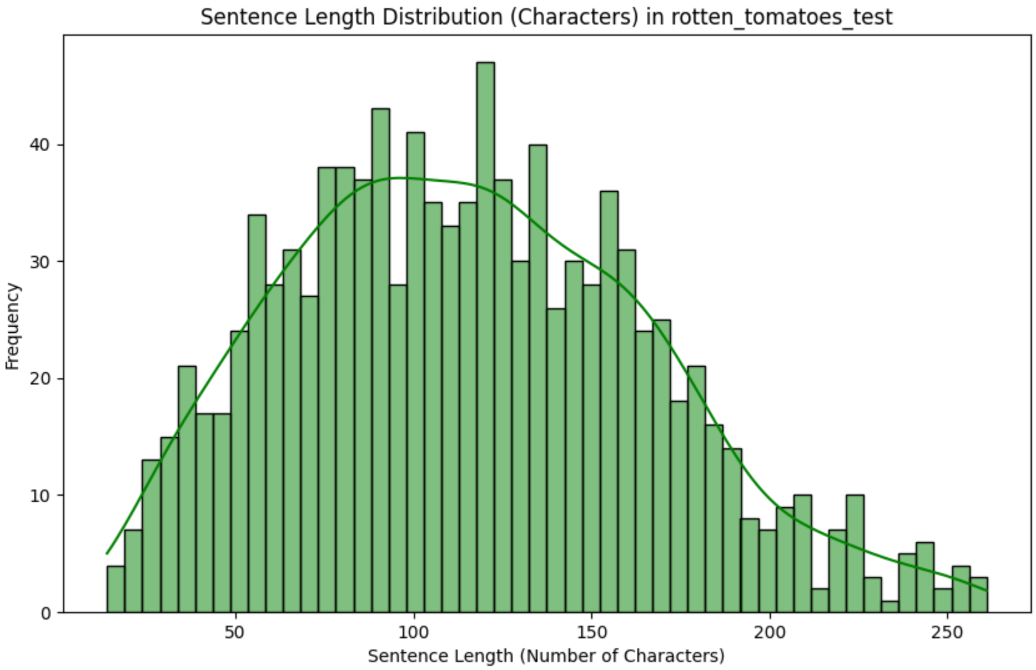
Label	Frequency
1	533
0	533



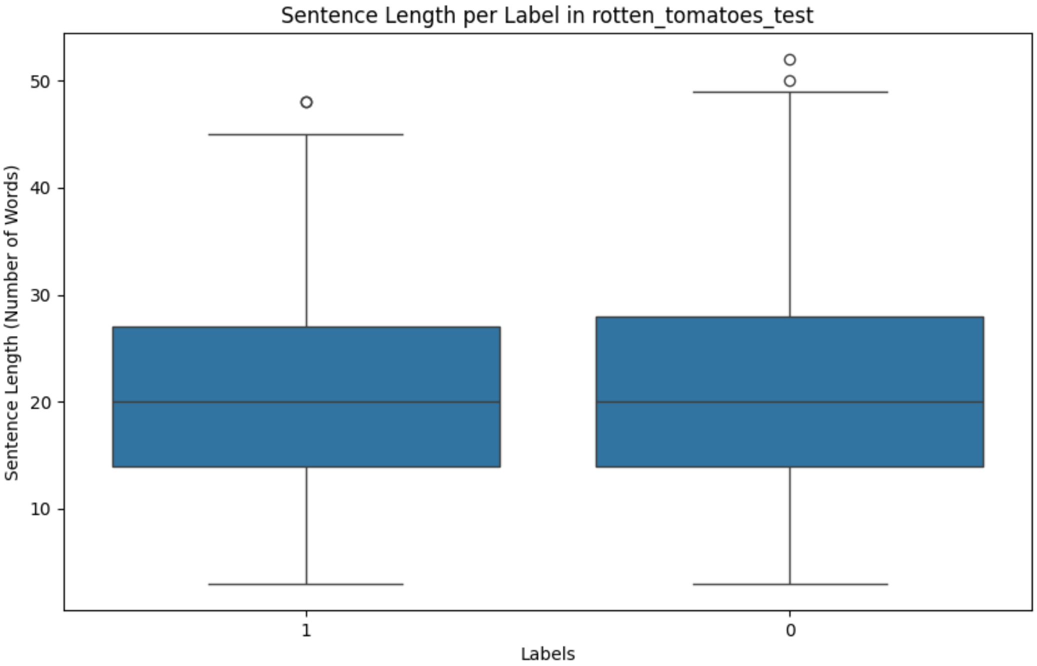
Sentence Length Distribution (Words)



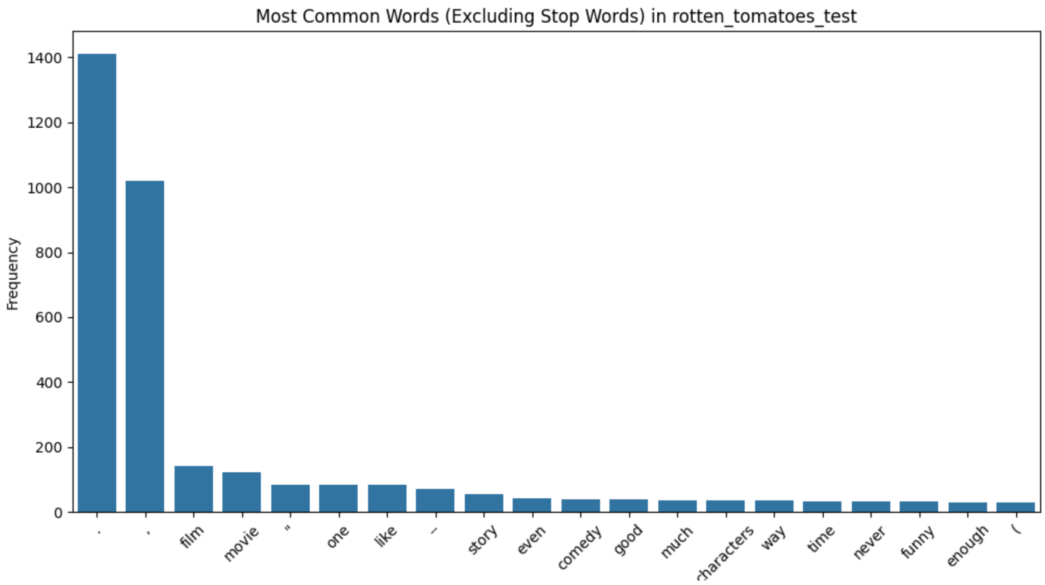
Sentence Length Distribution (Characters)



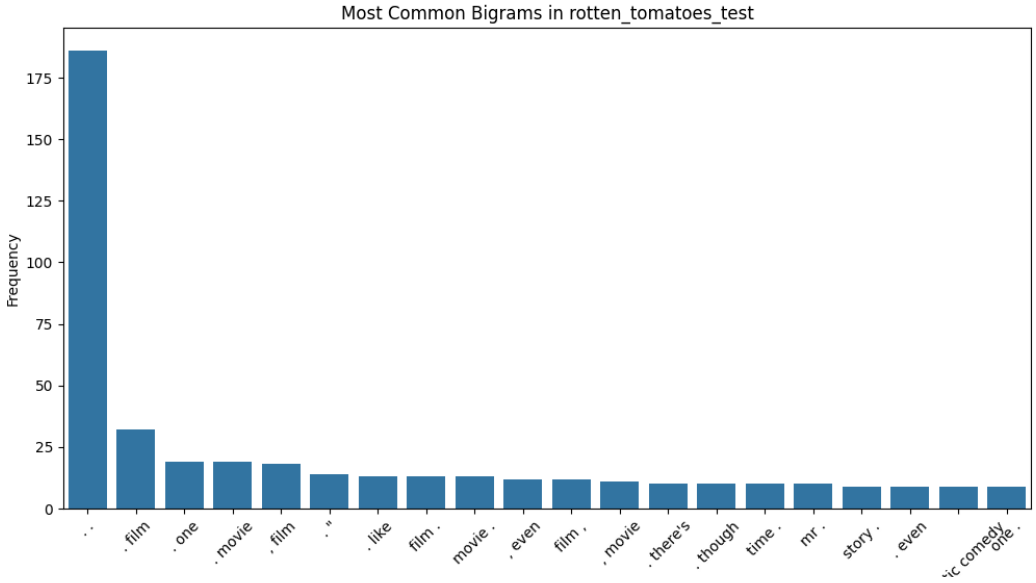
Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams



Analysis of rotten_tomatoes_train

Basic Information

- Number of samples: 8530
- Missing 'text' entries: 0
- Missing 'label' entries: 0

- Number of unique sentences: 8530
- Number of unique labels: 2
- Vocabulary size: 18951

Sentence Length (Words)

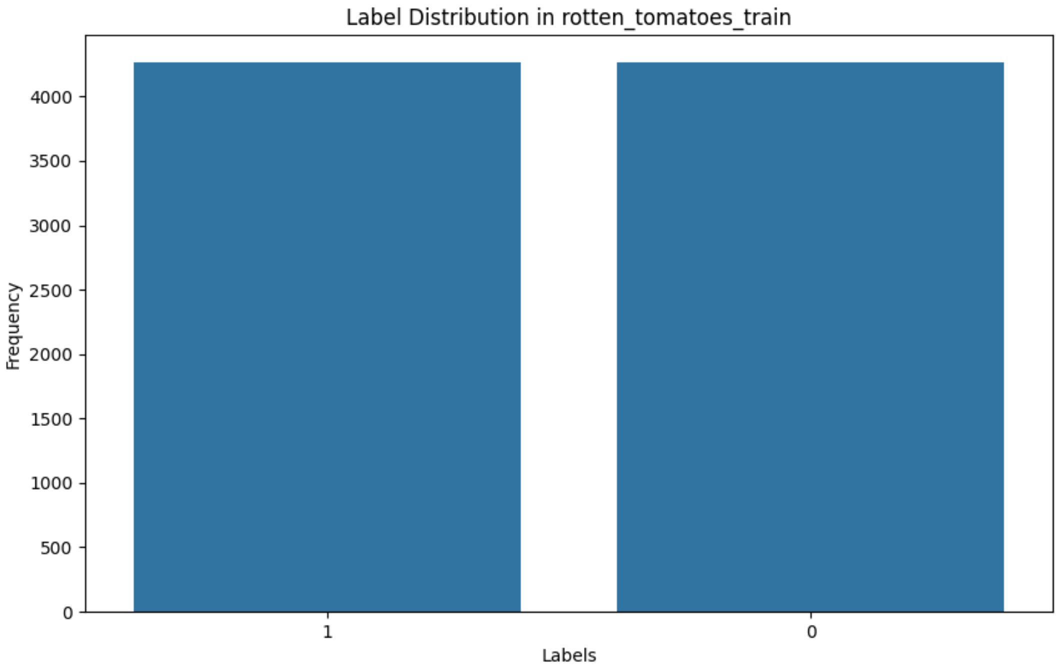
- Average: 20.99
- Standard deviation: 9.37
- Median: 20.0
- Max: 59
- Min: 1
- Quantiles (25%, 50%, 75%): 14.0, 20.0, 27.0

Stop Words Proportion

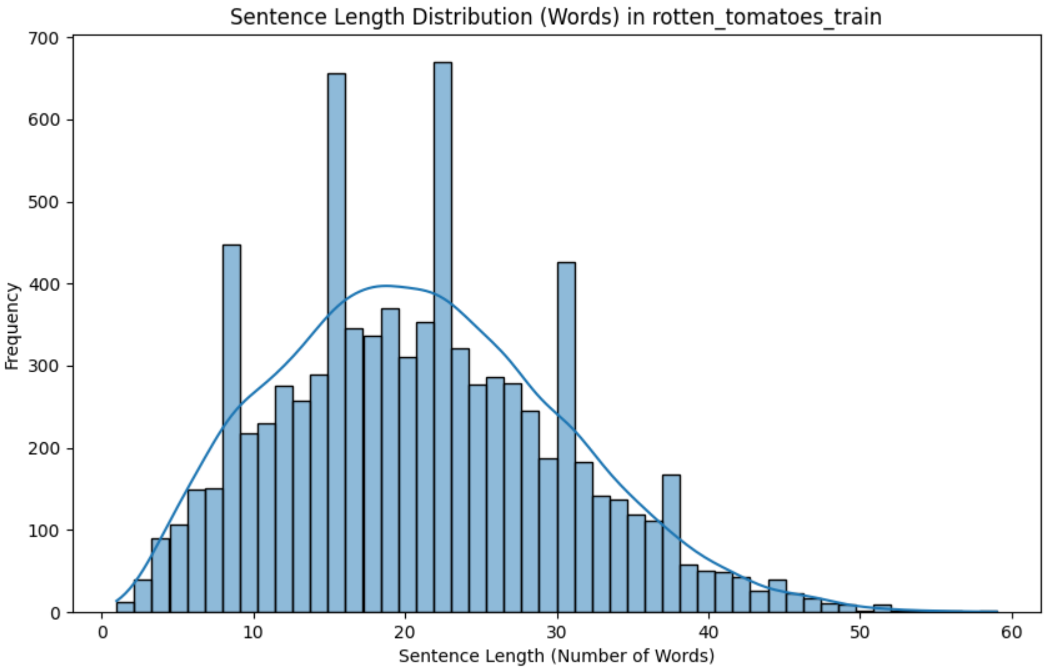
37.78%

Label Distribution

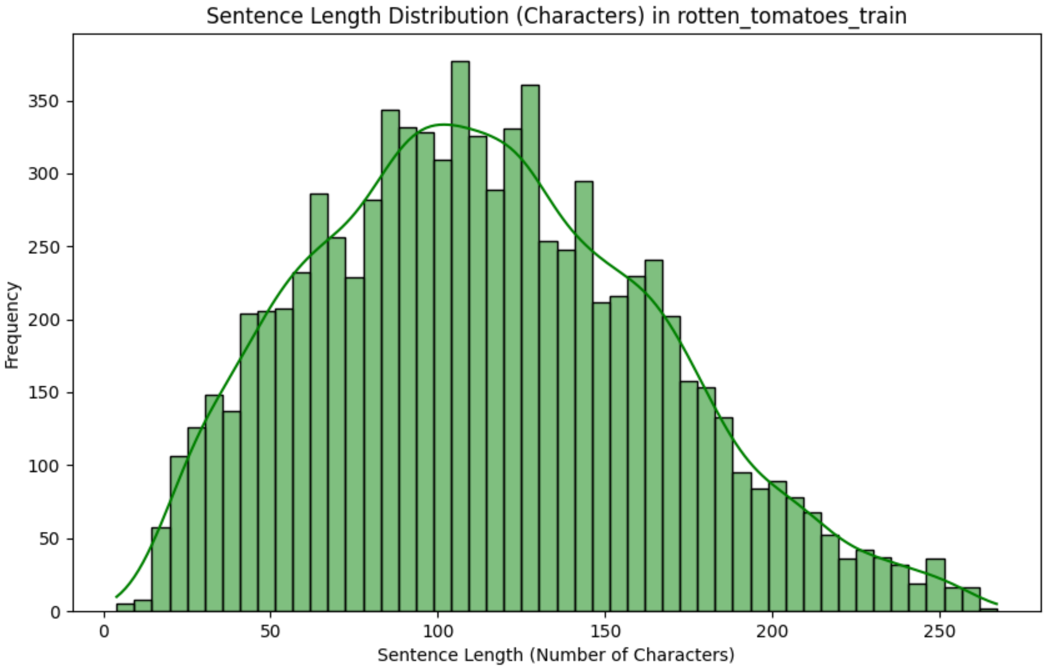
Label	Frequency
1	4265
0	4265



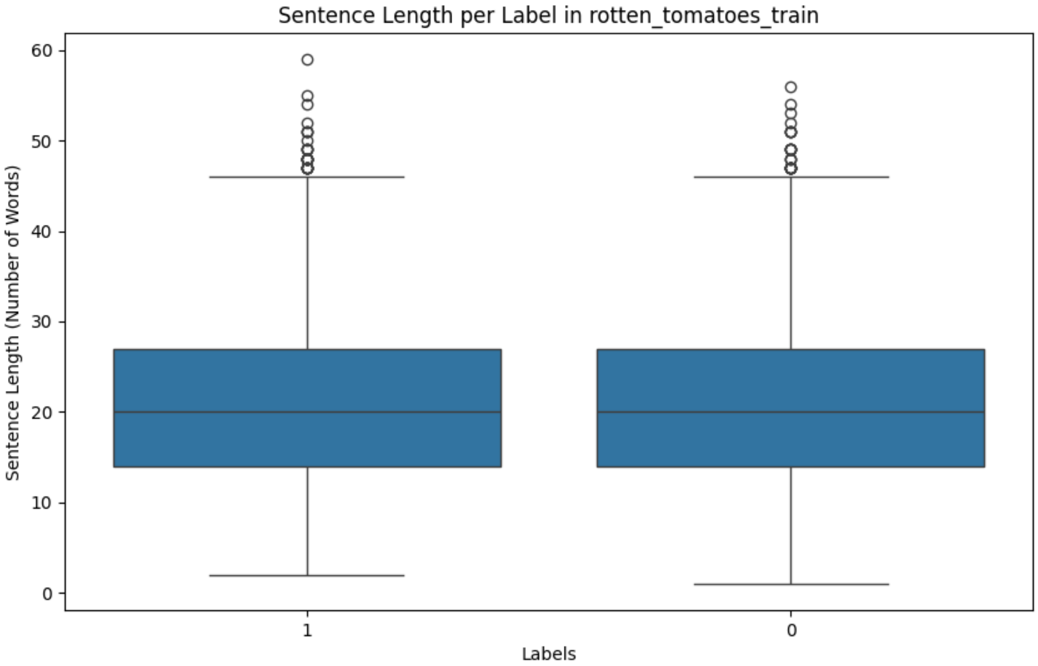
Sentence Length Distribution (Words)



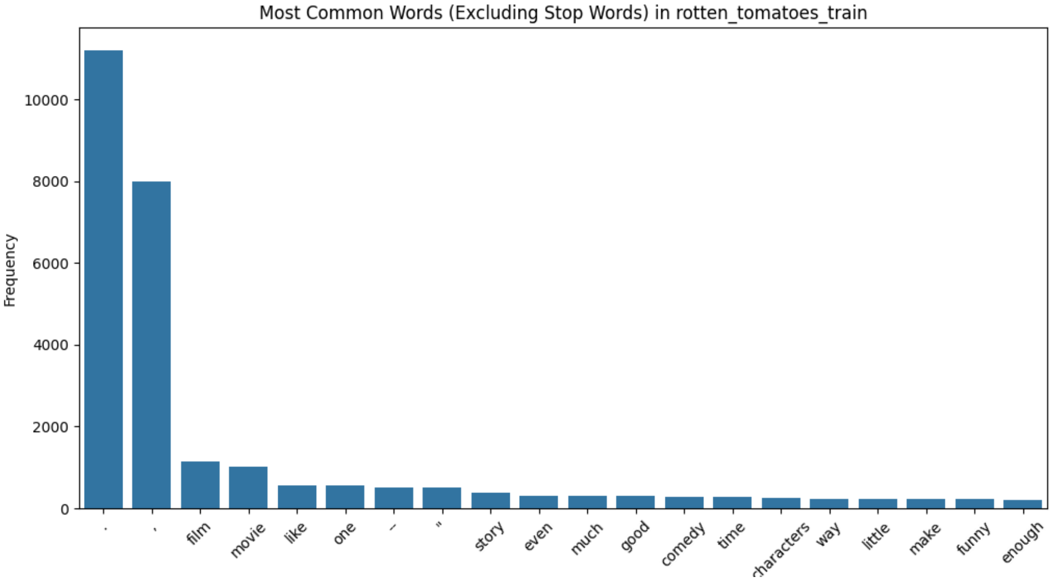
Sentence Length Distribution (Characters)



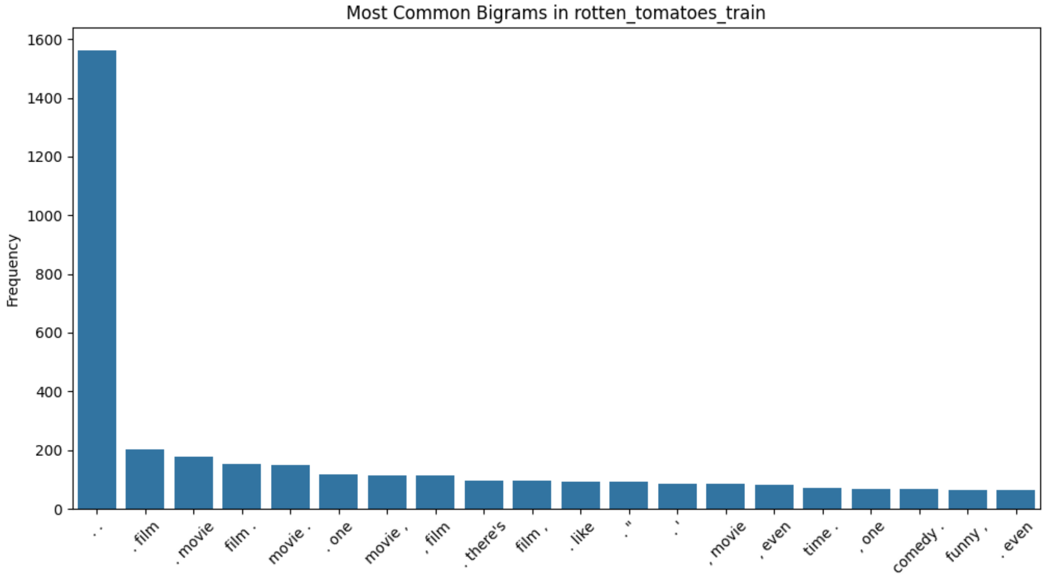
Sentence Length per Label



Most Common Words (Excluding Stop Words)



Most Common Bigrams


Most Common Bigrams in rotten_tomatoes_train
Frequency
. film
. movie
film .
movie .
. one
movie ,
, film
. there's
film ,
. like
. "
. '
, movie
, even
time .
, one
comedy .
funny ,
. even