MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Theory Activity No. 1

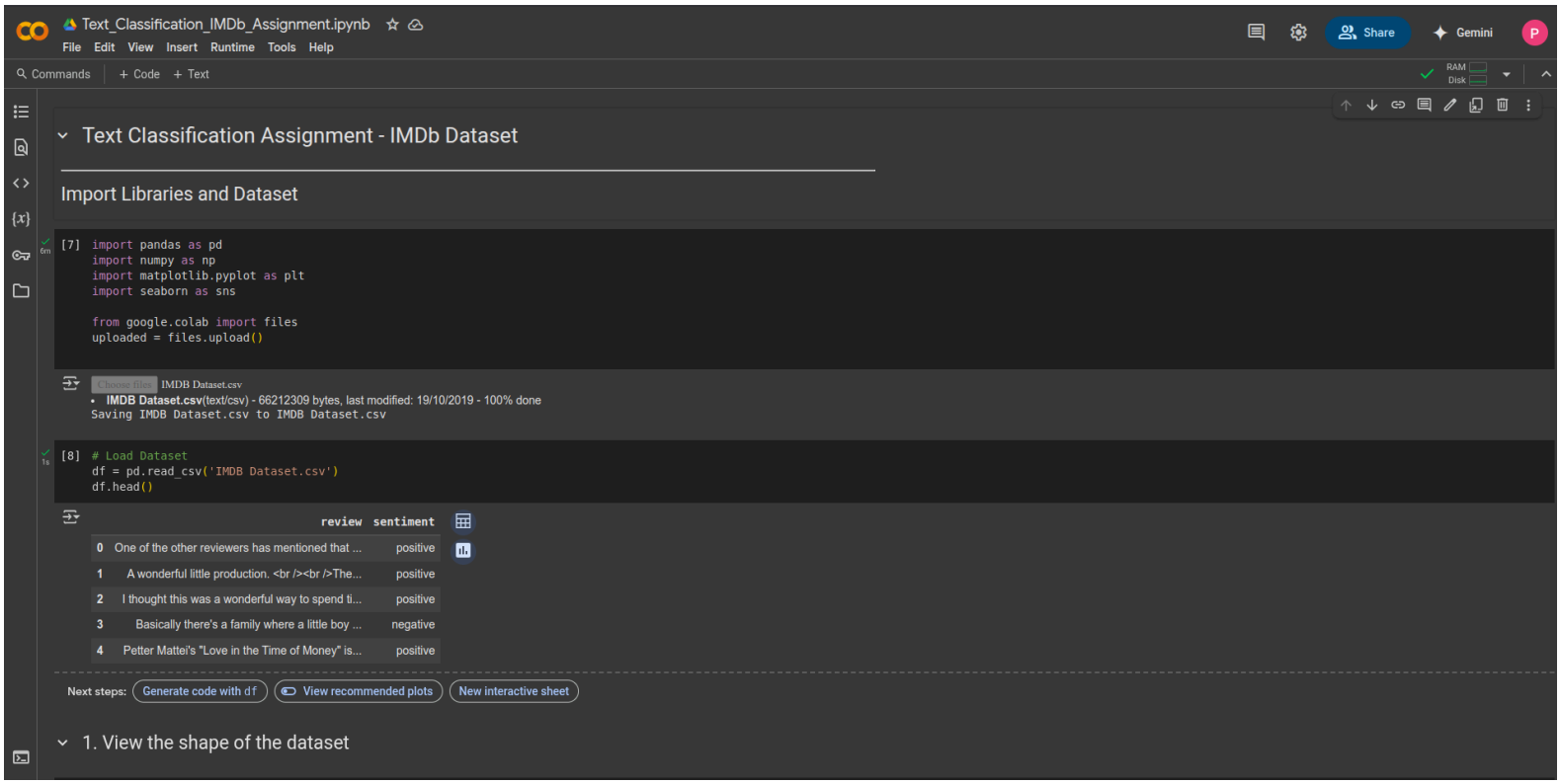
NAME: PUSHKAR HARSHKUMAR PARAKH

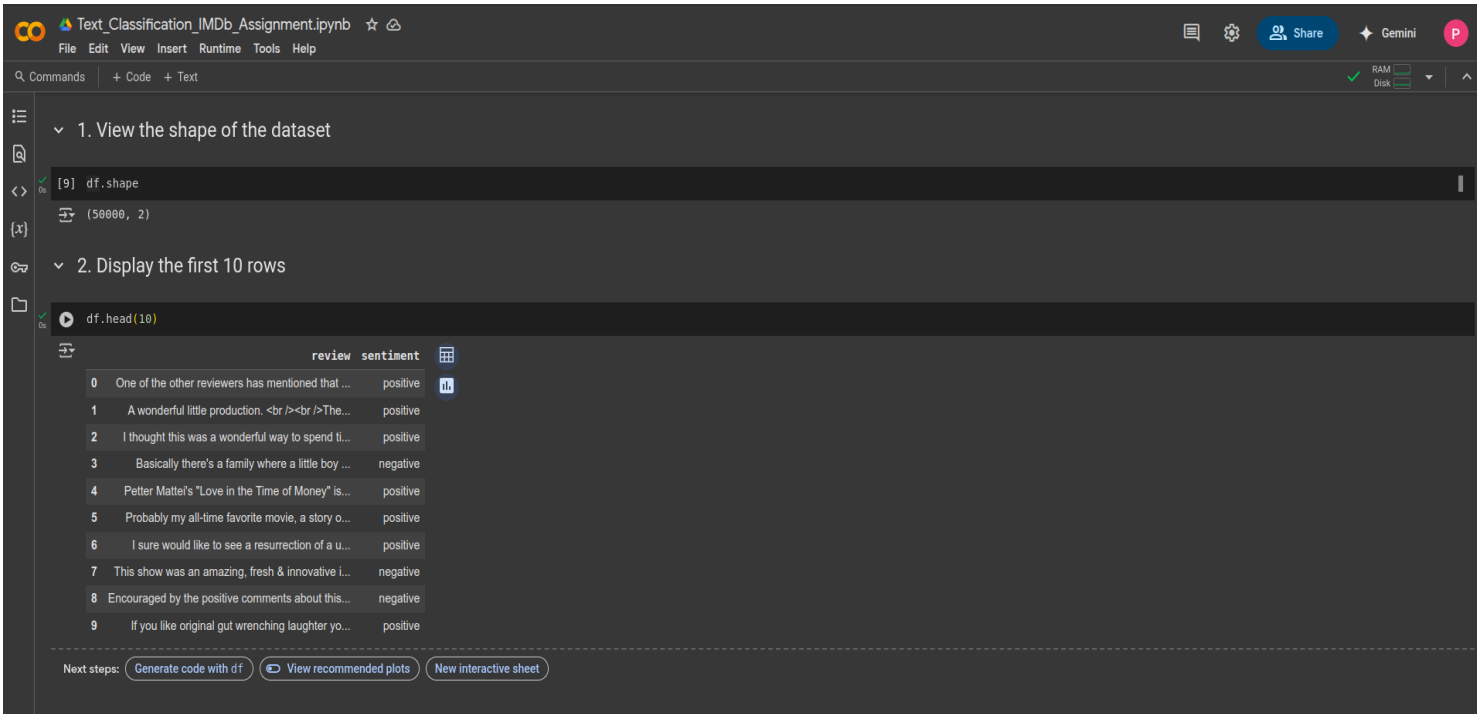
PRN: 202401040313

CLASS:CS4

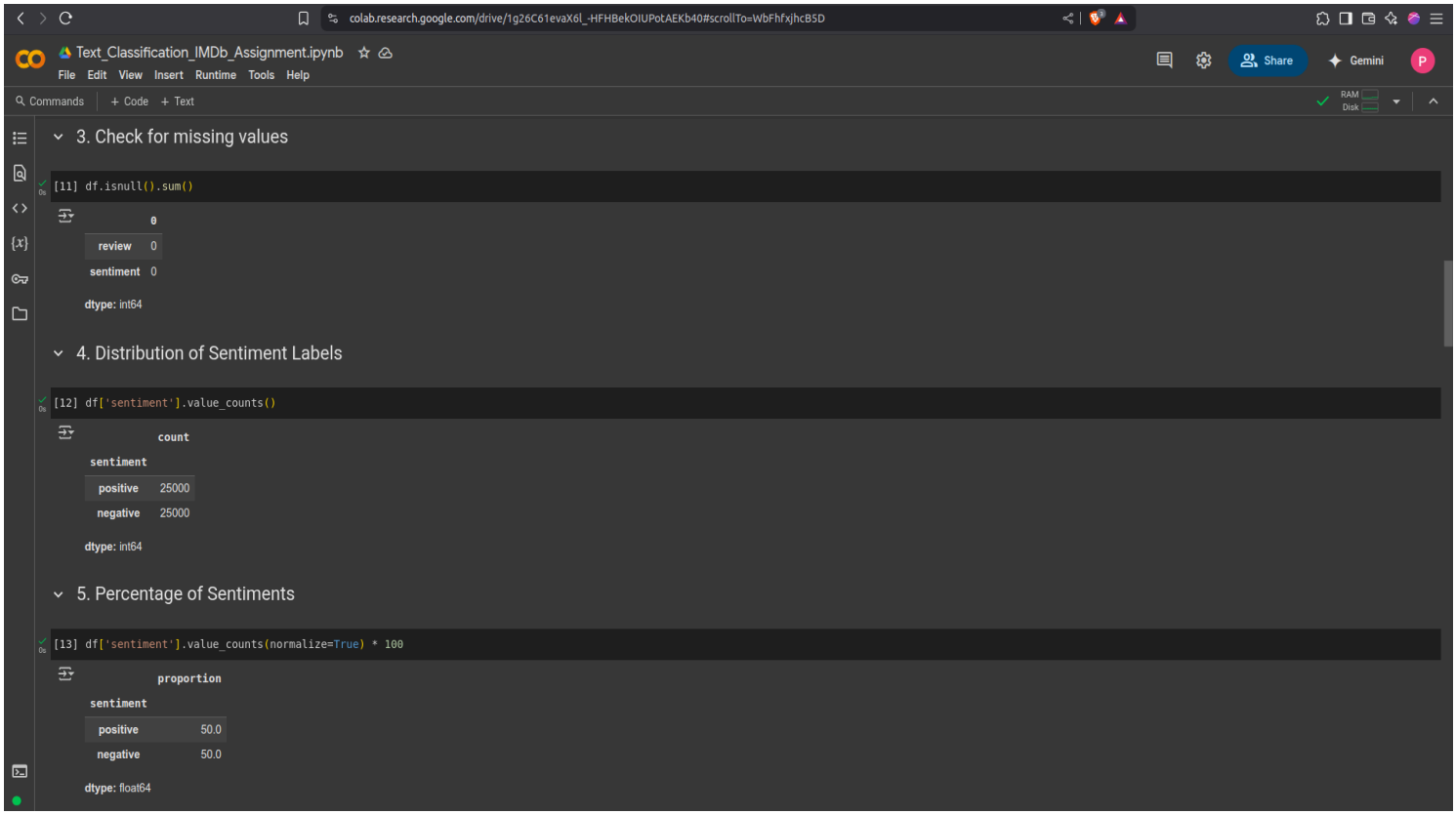
ROLL NO.:54

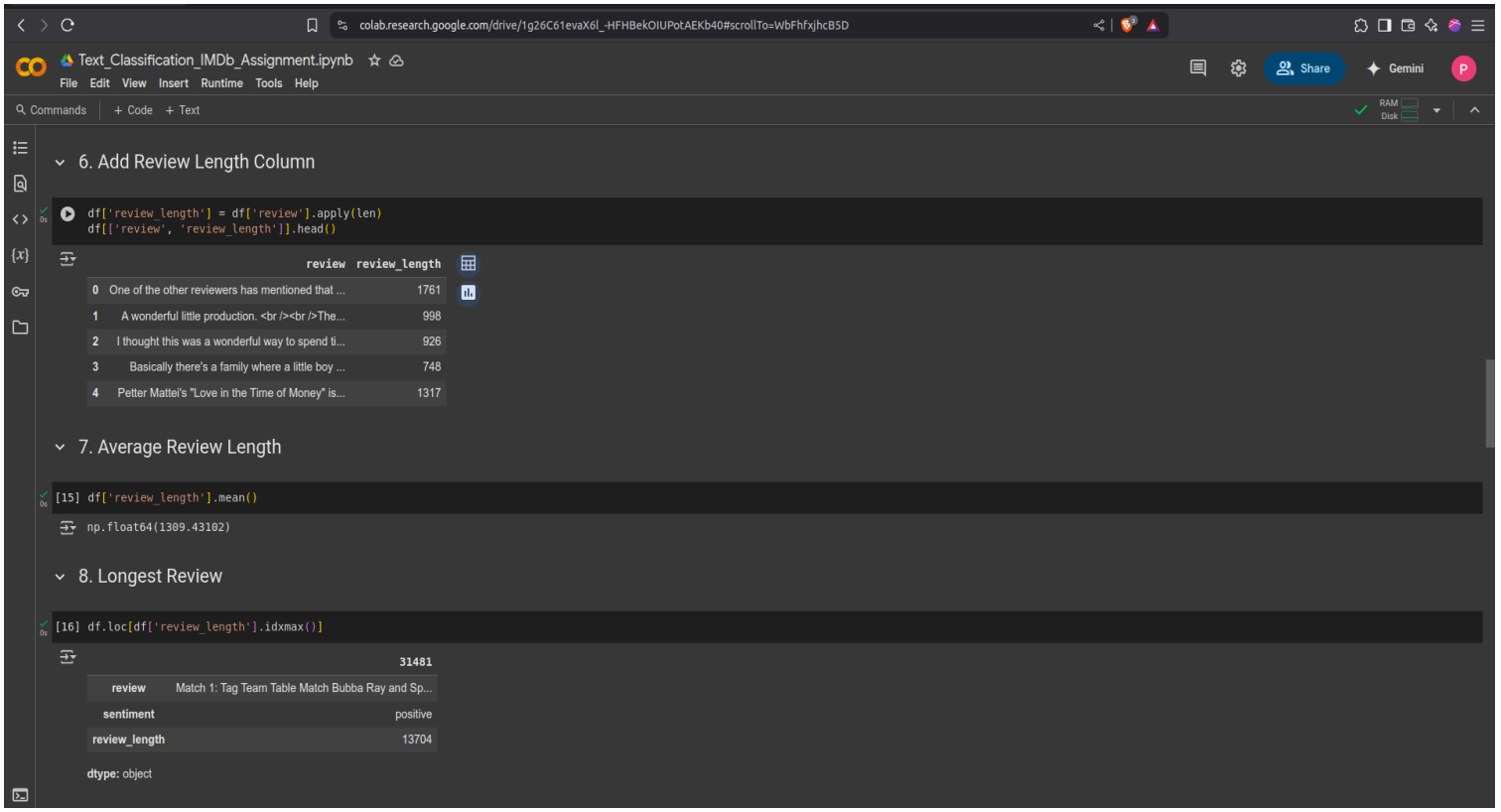
Text Classification Assignment - IMDb Dataset



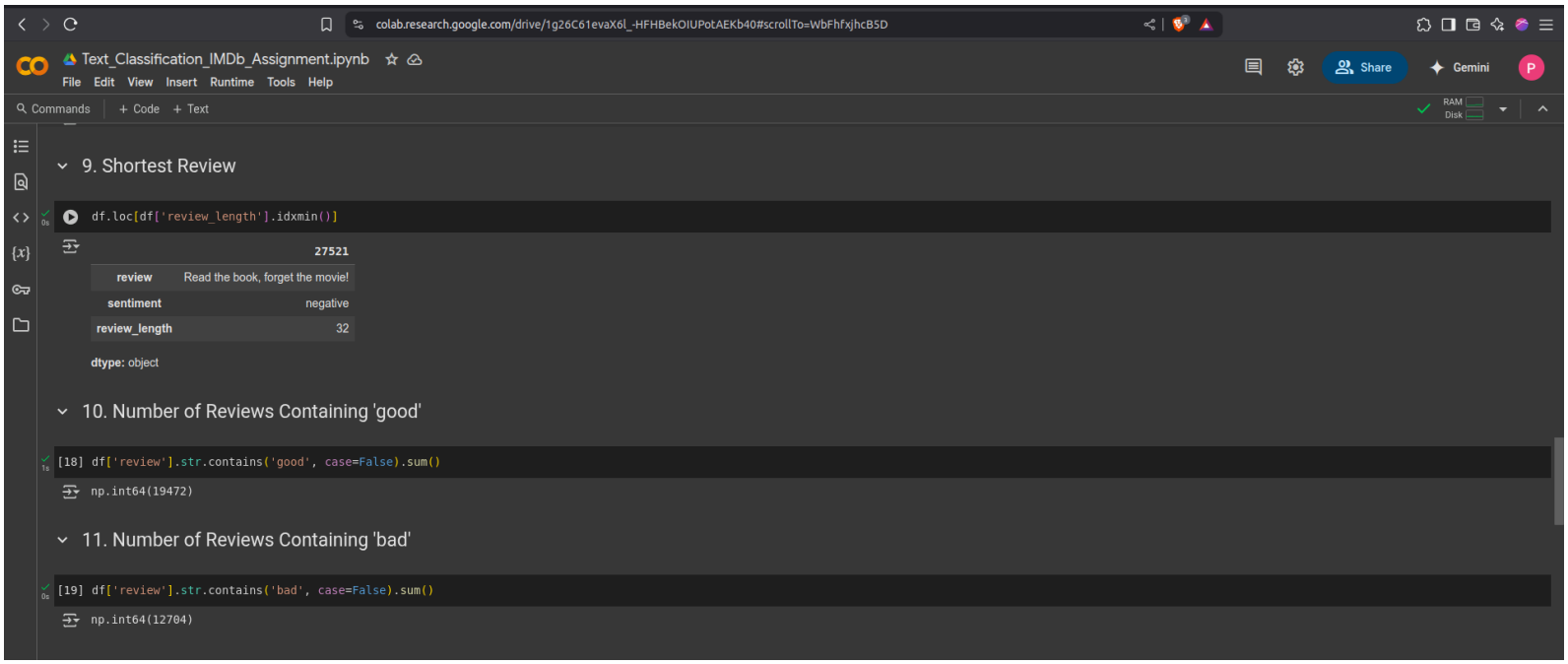

Text_Classification_IMDb_Assignment.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
1. View the shape of the dataset
[9] df.shape
(50000, 2)
2. Display the first 10 rows
df.head(10)
review sentiment
0 One of the other reviewers has mentioned that ... positive
1 A wonderful little production.

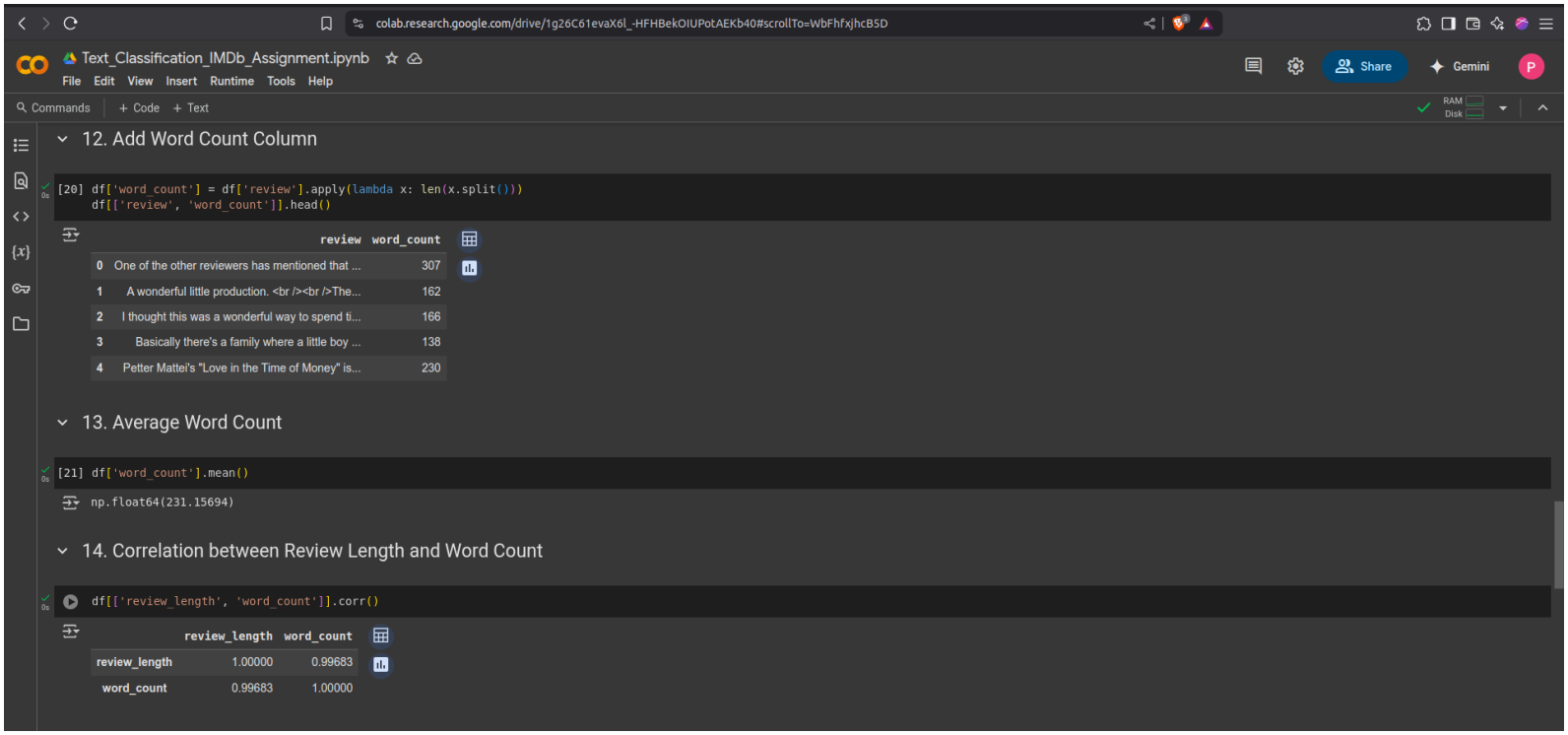
The... positive
2 I thought this was a wonderful way to spend ti... positive
3 Basically there's a family where a little boy ... negative
4 Petter Mattei's "Love in the Time of Money" is... positive
5 Probably my all-time favorite movie, a story o... positive
6 I sure would like to see a resurrection of a u... positive
7 This show was an amazing, fresh & innovative i... negative
8 Encouraged by the positive comments about this... negative
9 If you like original gut wrenching laughter yo... positive
Next steps:
Generate code with df
View recommended plots
New interactive sheet


colab.research.google.com/drive/1g26C61evaX6l_-HFHBekOIUPotAEKb40#scrollTo=WbFhfxjhcB5D
Text_Classification_IMDb_Assignment.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
3. Check for missing values
[11] df.isnull().sum()
0
review 0
sentiment 0
dtype: int64
4. Distribution of Sentiment Labels
[12] df['sentiment'].value_counts()
count
sentiment
positive 25000
negative 25000
dtype: int64
5. Percentage of Sentiments
[13] df['sentiment'].value_counts(normalize=True) * 100
proportion
sentiment
positive 50.0
negative 50.0
dtype: float64

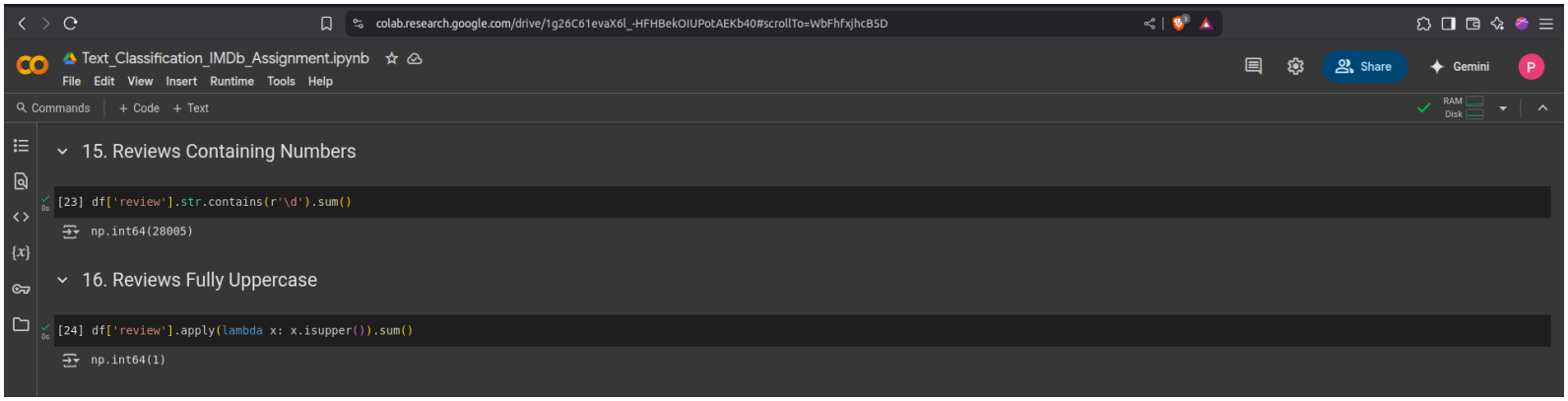

colab.research.google.com/drive/1g26C61evaX6l_-HFHBekOIUPotAEKb40#scrollTo=WbFhfxjhcB5D
Text_Classification_IMDb_Assignment.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
6. Add Review Length Column
df['review_length'] = df['review'].apply(len)
df[['review', 'review_length']].head()
review review_length
0 One of the other reviewers has mentioned that ... 1761
1 A wonderful little production.

The... 998
2 I thought this was a wonderful way to spend ti... 926
3 Basically there's a family where a little boy ... 748
4 Petter Mattei's "Love in the Time of Money" is... 1317
7. Average Review Length
[15] df['review_length'].mean()
np.float64(1309.43102)
8. Longest Review
[16] df.loc[df['review_length'].idxmax()]
31481
review Match 1: Tag Team Table Match Bubba Ray and Sp...
sentiment positive
review_length 13704
dtype: object


colab.research.google.com/drive/1g26C61evaX6l_-HFHBekOIUPotAEKb40#scrollTo=WbFhfxjhcB5D
Text_Classification_IMDb_Assignment.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
9. Shortest Review
df.loc[df['review_length'].idxmin()]
27521
review Read the book, forget the movie!
sentiment negative
review_length 32
dtype: object
10. Number of Reviews Containing 'good'
[18] df['review'].str.contains('good', case=False).sum()
np.int64(19472)
11. Number of Reviews Containing 'bad'
[19] df['review'].str.contains('bad', case=False).sum()
np.int64(12704)


colab.research.google.com/drive/1g26C61evaX6l_-HFHBekOIUPotAEKb40#scrollTo=WbFhfxjhcB5D
Text_Classification_IMDb_Assignment.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
12. Add Word Count Column
[20] df['word_count'] = df['review'].apply(lambda x: len(x.split()))
df[['review', 'word_count']].head()
review word_count
0 One of the other reviewers has mentioned that ... 307
1 A wonderful little production.

The... 162
2 I thought this was a wonderful way to spend ti... 166
3 Basically there's a family where a little boy ... 138
4 Petter Mattei's "Love in the Time of Money" is... 230
13. Average Word Count
[21] df['word_count'].mean()
np.float64(231.15694)
14. Correlation between Review Length and Word Count
df[['review_length', 'word_count']].corr()
review_length word_count
review_length 1.00000 0.99683
word_count 0.99683 1.00000


colab.research.google.com/drive/1g26C61evaX6l_-HFHBekOIUPotAEKb40#scrollTo=WbFhfxjhcB5D
Text_Classification_IMDb_Assignment.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
15. Reviews Containing Numbers
[23] df['review'].str.contains(r'\d').sum()
np.int64(28005)
16. Reviews Fully Uppercase
[24] df['review'].apply(lambda x: x.isupper()).sum()
np.int64(1)

colab.research.google.com/drive/1g26C61evaX6l_-HFHBekOIUPotAEKb40#scrollTo=WbFhfxjhcB5D

Text_Classification_IMDb_Assignment.ipynb

File Edit View Insert Runtime Tools Help

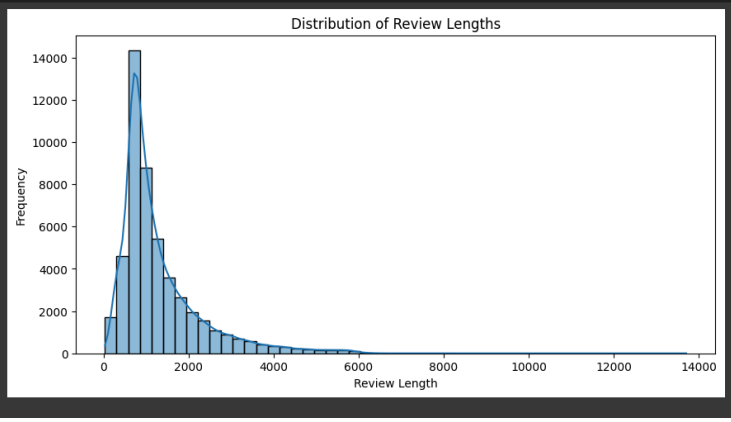
Share Gemini

Commands + Code + Text

RAM Disk

17. Distribution of Review Lengths

```
plt.figure(figsize=(10,5))
sns.histplot(df['review_length'], bins=50, kde=True)
plt.title('Distribution of Review Lengths')
plt.xlabel('Review Length')
plt.ylabel('Frequency')
plt.show()
```



colab.research.google.com/drive/1g26C61evaX6l_-HFHBekOIUPotAEKb40#scrollTo=WbFhfxjhcB5D

Text_Classification_IMDb_Assignment.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands + Code + Text

RAM Disk

18. Average Review Length by Sentiment

```
[26] df.groupby('sentiment')['review_length'].mean()
```

sentiment	review_length
negative	1294.06436
positive	1324.79768

dtype: float64

19. Review with Maximum Words

```
df.loc[df['word_count'].idxmax()]
```

	31481
review	Match 1: Tag Team Table Match Bubba Ray and Sp...
sentiment	positive
review_length	13704
word_count	2470

dtype: object

20. Top 10 Common Words

```
[28] from collections import Counter
     import re

     all_text = ' '.join(df['review']).lower()
     words = re.findall(r'\b\w+\b', all_text)
     word_counts = Counter(words)
     word_counts.most_common(10)
```

```
[('the', 667993),
 ('and', 324441),
 ('a', 322970),
 ('of', 289410),
 ('to', 268124),
 ('is', 211082),
 ('br', 201951),
 ('it', 190857),
 ('in', 186781),
 ('i', 175633)]
```