02_eda_raw

October 12, 2025

1 Step 2 — EDA on Raw Data

Goal: Understand the freshly scraped Reddit data before heavy cleaning.

What this notebook does

- Loads `pipeline dataset` from the scraping step.
- Basic checks: row count, missing values, duplicates, class/topic balance, and time coverage.
- Text previews: average length, token counts, frequent words, simple n-grams.
- Quality flags: identifies spammy/very short/very long entries for later filtering.

Inputs: `pipeline dataset`.

[1]:

```
try:
    IS_PIPELINE_RUN
except NameError:
    IS_PIPELINE_RUN = False

try:
    IS_PIPELINE_TEST
except NameError:
    IS_PIPELINE_TEST = False
```

[1]:

```
import pandas as pd
import matplotlib.pyplot as plt


from pathlib import Path
```

[ ]:

```
dataset_folder = Path("../datasets")

if IS_PIPELINE_RUN:
    filename1 = "Palo_Alto_pipeline_reddit.pkl"
else:
    filename1 = "Palo_Alto_20251007_235943_reddit.pkl"

# read palo alto dataframe
df = pd.read_pickle(dataset_folder / filename1)
df.head()
```

[ ]:

```
   source                                               query  \
0  reddit  "Palo Alto" (school OR schools OR district OR …
1  reddit  "Palo Alto" (school OR schools OR district OR …
2  reddit  "Palo Alto" (school OR schools OR district OR …
3  reddit  "Palo Alto" (school OR schools OR district OR …
4  reddit  "Palo Alto" (school OR schools OR district OR …

                                                topic  \
0  Mark Zuckerberg and his wife shut down their s…
1  Mark Zuckerberg and his wife shut down their s…
2  Bay Area teen rejected by 16 colleges, hired b…
3  To be a philanthropist. Mark Zuckerberg and hi…
4  Are you kicking kids out by 18? (Or if you wer…

                                     comments_nested  \
0  [[It's almost as if schools should be funded b…
1  [[Delete your Facebook.  Delete IG.  Delete Wh…
2  [[Quote from another comment on this topic fro…
3  [[Nothing says "cutting-edge wave of the futur…
4  [[Fuck no. I am doing my best to ensure that t…

                                       comments_flat  num_comments  \
0  [It's almost as if schools should be funded by…              84
1  [Delete your Facebook.  Delete IG.  Delete Wha…              21
2  [Quote from another comment on this topic from…             128
3  [Nothing says "cutting-edge wave of the future…              10
4  [Fuck no. I am doing my best to ensure that th…             157

   total_words
0         2722
1          560
2         6529
3          313
4         9365
```

2 Palo Alto EDA Raw Data

[4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 79 entries, 0 to 78
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   source          79 non-null     object
 1   query           79 non-null     object
 2   topic           79 non-null     object
 3   comments_nested  79 non-null     object
 4   comments_flat    79 non-null     object
 5   num_comments     79 non-null     int64
 6   total_words      79 non-null     int64
dtypes: int64(2), object(5)
memory usage: 4.4+ KB
```

[5]:

```
cat_cols = ["source", "query", "topic", "comments_nested", "comments_flat"]
num_cols = ["num_comments", "total_words"]
```

[6]:

```
df[cat_cols].describe(include="all")
```

[6]:

```
        source                                              query  \
count       79                                                 79
unique       1                                                  1
top     reddit  "Palo Alto" (school OR schools OR district OR …
freq        79                                                 79

                                                topic  \
count                                              79
unique                                             74
top     Mark Zuckerberg and his wife shut down their s…
freq                                                4

                                      comments_nested  \
count                                              79
unique                                             79
top     [[It's almost as if schools should be funded b…
freq                                                1

                                        comments_flat
count                                              79
unique                                             79
top     [It's almost as if schools should be funded by…
freq                                                1
```

[7]:

```
df[num_cols].describe(include="all")
```

[7]:

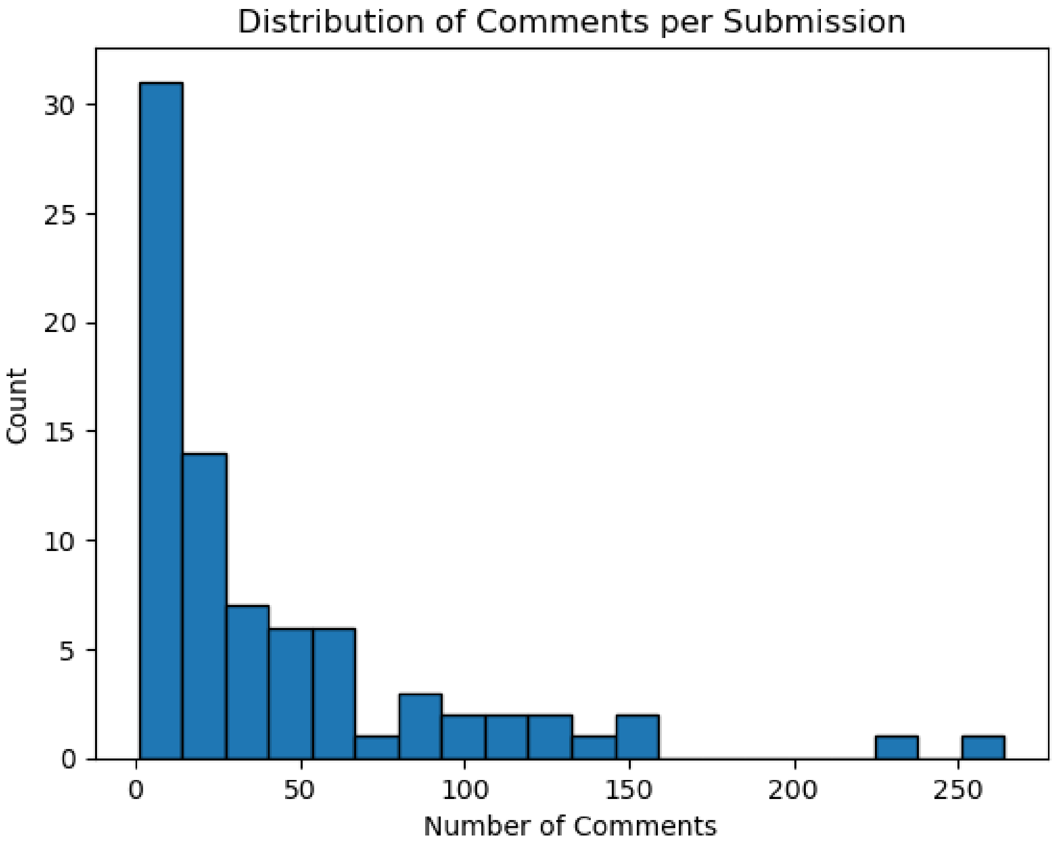
```
       num_comments   total_words
count     79.000000     79.000000
mean      41.873418   2179.962025
std       51.706858   3140.477007
min        1.000000     21.000000
25%        7.000000    403.000000
50%       21.000000    951.000000
75%       55.500000   2635.500000
max      264.000000  18503.000000
```

[8]:

```
def show_histogram(
    dataframe, column, title=None, xlabel=None, bins=20, edgecolor="black"
):
    plt.hist(dataframe[column], bins=bins, edgecolor=edgecolor)
    plt.title(title or "Distribution")
    plt.xlabel(xlabel or column)
    plt.ylabel("Count")
    plt.show()
```

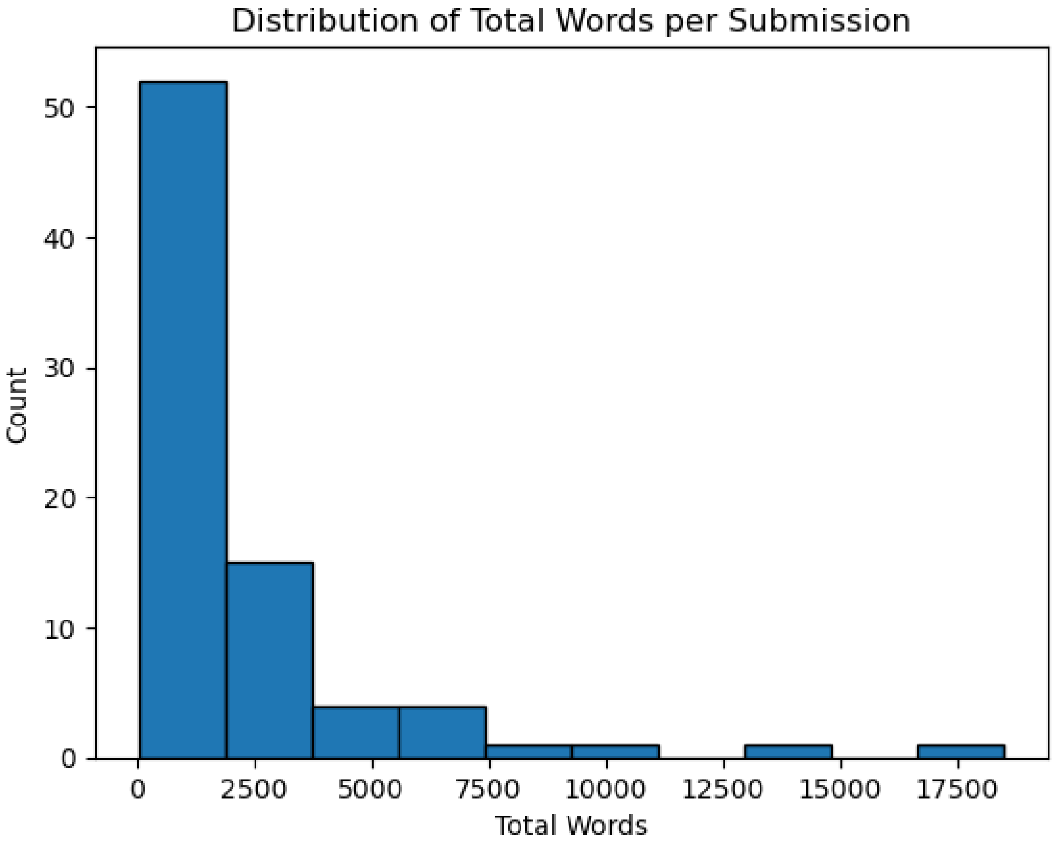
[9]:

```
show_histogram(
    df,
    "num_comments",
    title="Distribution of Comments per Submission",
    xlabel="Number of Comments",
)
```



[10]:

```
show_histogram(
    df,
    "total_words",
    title="Distribution of Total Words per Submission",
    xlabel="Total Words",
    bins=10,
)
```



3 Oklahoma City EDA Raw Data

[ ]:

```
# read ok city dataframe
if IS_PIPELINE_RUN:
    filename2 = "Oklahoma_City_pipeline_reddit.pkl"
else:
    filename2 = "Oklahoma_City_20251008_000300_reddit.pkl"
df1 = pd.read_pickle(dataset_folder / filename2)

df1.head()
```

[ ]:

```
   source                                              query  \
0  reddit  "Oklahoma City" (school OR schools OR district…
1  reddit  "Oklahoma City" (school OR schools OR district…
2  reddit  "Oklahoma City" (school OR schools OR district…
3  reddit  "Oklahoma City" (school OR schools OR district…
4  reddit  "Oklahoma City" (school OR schools OR district…

                                               topic  \
0  6 Oklahoma City teachers fired for refusing to…
1  Russell Westbrook visited Oklahoma City today …
2  Jalen Williams on the Oklahoma City crowd: "It…
3  Russell Westbrook has been named as the Creati…
4  U.S. Counties Where the Non-Hispanic White Ame…

                                     comments_nested  \
0  [[I assure you that if these teachers refused …
1  [[Russ will forever be a legend here and has d…
2  [[Isaiah Stewart on his way to demand a trade …
3  [[To be clear, this is NOT regarding the new O…
4  [[I knew it all along but there is a sizeable …

                                       comments_flat  num_comments  \
0  [I assure you that if these teachers refused t…           105
1  [Russ will forever be a legend here and has do…           100
2  [Isaiah Stewart on his way to demand a trade t…            36
3  [To be clear, this is NOT regarding the new OK…            18
4  [I knew it all along but there is a sizeable H…           177

   total_words
0         3385
1         3658
2          986
3          482
4         6329
```

[13]:

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 135 entries, 0 to 134
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   source           135 non-null    object
 1   query            135 non-null    object
 2   topic            135 non-null    object
 3   comments_nested  135 non-null    object
 4   comments_flat    135 non-null    object
 5   num_comments     135 non-null    int64
 6   total_words      135 non-null    int64
dtypes: int64(2), object(5)
memory usage: 7.5+ KB
```

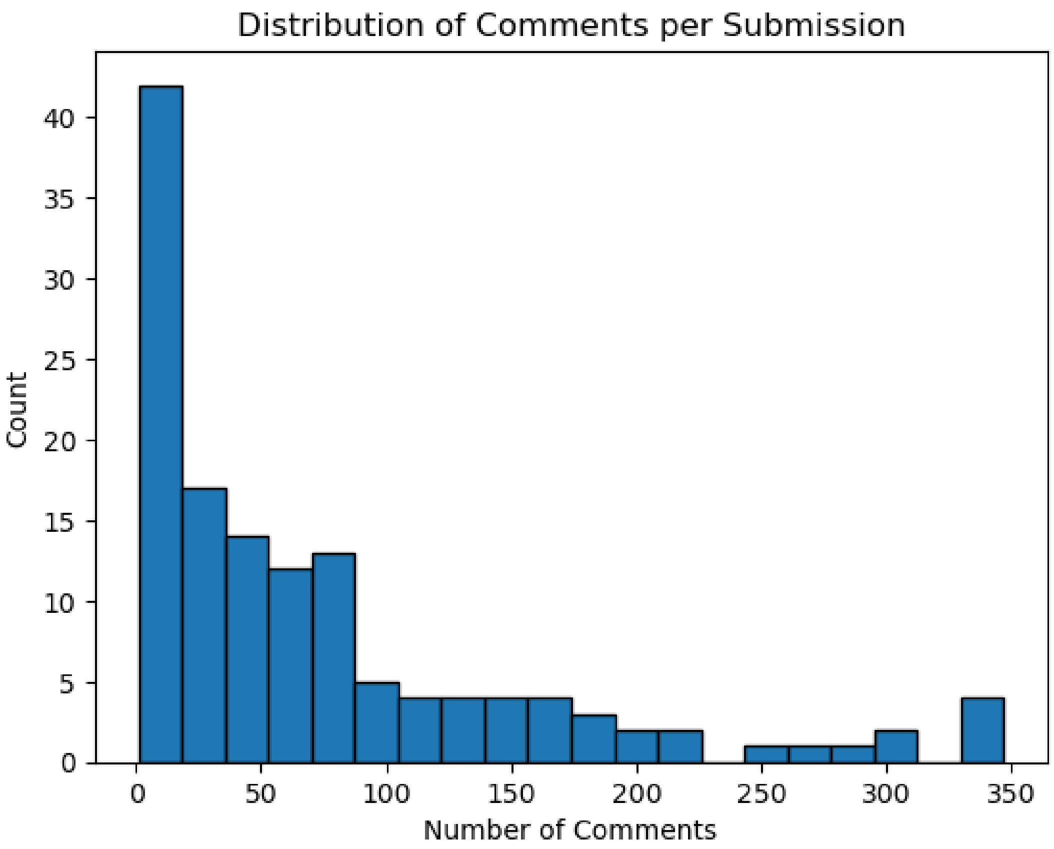
[14]:
```
df1[cat_cols].describe(include="all")
```

[14]:
```
        source                                              query  \
count      135                                                135
unique       1                                                  1
top     reddit  "Oklahoma City" (school OR schools OR district…
freq       135                                                135

                                                    topic  \
count                                                 135
unique                                                131
top     6 Oklahoma City teachers fired for refusing to…
freq                                                    2

                                          comments_nested  \
count                                                 135
unique                                                135
top     [[I assure you that if these teachers refused …
freq                                                    1

                                            comments_flat
count                                                 135
unique                                                135
top     [I assure you that if these teachers refused t…
freq                                                    1
```

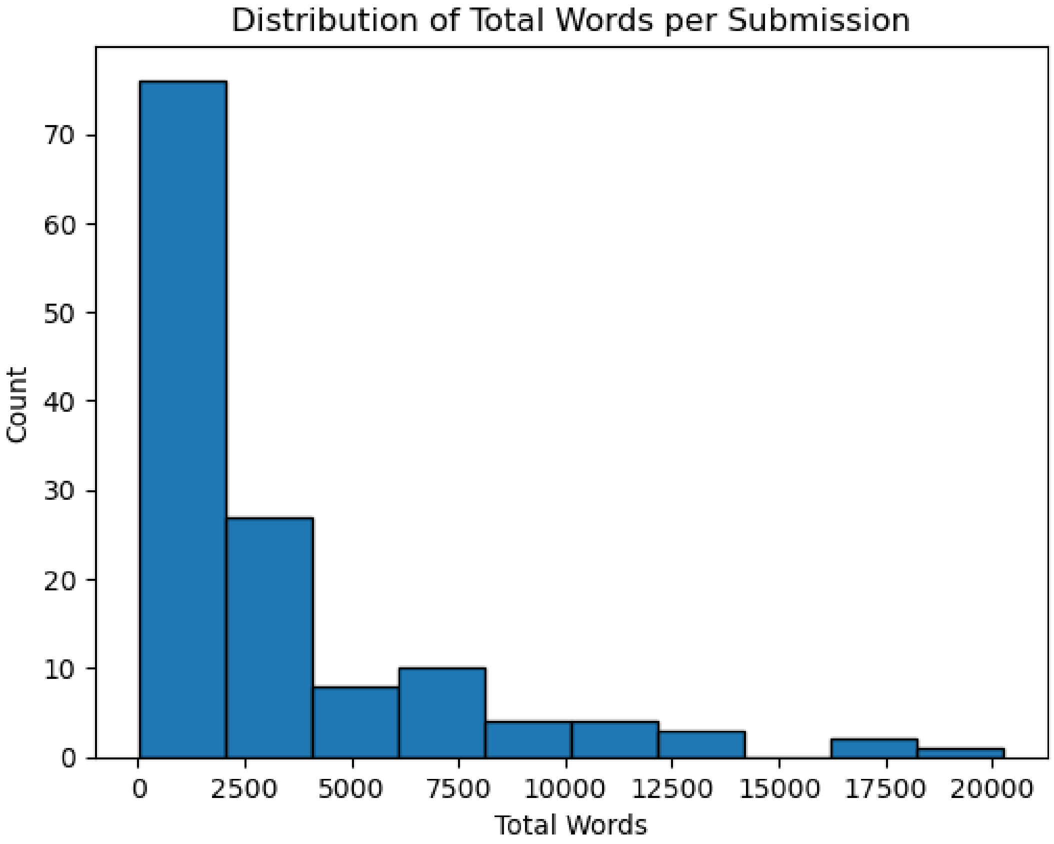
[15]:
```
df1[num_cols].describe(include="all")
```

[15]:
```
       num_comments   total_words
count    135.000000    135.000000
mean      74.148148   3175.459259
std       82.922596   3902.108983
min        1.000000     15.000000
25%       14.000000    570.500000
50%       43.000000   1526.000000
75%       98.500000   3971.500000
max      347.000000  20272.000000
```

[16]:
```
show_histogram(
    df1,
    "num_comments",
    title="Distribution of Comments per Submission",
    xlabel="Number of Comments",
)
```



[17]:
```
show_histogram(
    df1,
    "total_words",
    title="Distribution of Total Words per Submission",
    xlabel="Total Words",
    bins=10,
)
```


Distribution of Total Words per Submission
Count
0
10
20
30
40
50
60
70
0
2500
5000
7500
10000
12500
15000
17500
20000
Total Words