Analyzing Comment Structures for Popularity Prediction

In this project, we conducted an in-depth analysis of user-generated comments on a social media platform to better understand the factors influencing comment popularity. The analysis involved investigating the linguistic features and patterns within comments posted in response to various posts. Here are the key findings:

In [ ]:
```
# Data loading. Data description
```

In [2]:
```
import pandas as pd
import json
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
def from_json_to_df(path: str):
    data = []

    with open(path, 'r') as f:
        for line in f:
            sample = json.loads(line)
            post_text = sample['text']
            for comment in sample['comments']:
                comment_text = comment['text']
                score = comment['score']
                data.append([post_text, comment_text, score])

    return pd.DataFrame(data, columns=['post_text', 'comment_text', 'score'])
```

In [4]:
```
train_df = from_json_to_df("C:\\Users\\marty\\Desktop\\data\\ranking_train.json")
print(train_df.shape)
    train_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 446535 entries, 0 to 446534
Data columns (total 3 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   post_text     446535 non-null  object
 1   comment_text  446535 non-null  object
 2   score         446535 non-null  int64
dtypes: int64(1), object(2)
memory usage: 10.2+ MB
(446535, 3) None
```

In [5]:
```
train_df.head(10)
```

Out[6]:

	post_text	comment_text	score
0	How many summer Y Combinator funders decided n...	Going back to school is not identical with giv...	0
1	How many summer Y Combinator funders decided n...	There will invariably be those who don't see t...	1
2	How many summer Y Combinator funders decided n...	For me school is a way to be connected to what...	2
3	How many summer Y Combinator funders decided n...	I guess it really depends on how hungry you ar...	3
4	How many summer Y Combinator funders decided n...	I know poligrounsd decided to go back to school...	4
5	CBS acquires last.fm for $280m	It will be curious to see where this heads in ...	0
6	CBS acquires last.fm for $280m	Does this mean that there's now a big-name com...	1
7	CBS acquires last.fm for $280m	Also on BBC News: http://news.bbc.co.uk/2/hi/e...	2
8	CBS acquires last.fm for $280m	I don't understand what they do that is worth ...	3
9	CBS acquires last.fm for $280m	sold out too cheaply. given their leadership p...	4

For each post, 5 comments were extracted, and each comment was assigned a popularity score: 0 for the most popular comment, and 4 for the least popular.

Primary analysis

Hypothesis 1: the larger the comment (the more words it contains), the more informative it is - the more popular it is

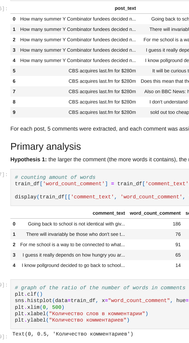
In [7]:
```
# counting amount of words
train_df['word_count_comment'] = train_df['comment_text'].apply(lambda x : len(str(x).split(" ")))

display(train_df[['comment_text', 'word_count_comment', 'score']].head(1))
```

	comment_text	word_count_comment	score
0	Going back to school is not identical with giv...	186	0

In [9]:
```
# graph of the ratio of the number of words in comments
plt.clf()
sns.histplot(datartrain_df, x="word_count_comment", hue="score", bins=1000)
plt.xlim(0, 500)
plt.xlabel("Количество слов в комментарии")
plt.ylabel("Количество комментариев")
```

Out[9]:
```
Text(0, 0.5, 'Количество комментариев')
```



In [10]:
```
target_0 = train_df.loc[train_df['score'] == 0]
target_1 = train_df.loc[train_df['score'] == 1]
target_2 = train_df.loc[train_df['score'] == 2]
target_3 = train_df.loc[train_df['score'] == 3]
target_4 = train_df.loc[train_df['score'] == 4]
```

In [11]:
```
print("Среднее количество слов в самом популярном комментарии (0 score) -", target_0['word_count_comment'].mean() )
print("Медианное количество слов в самом популярном комментарии (0 score) -", target_0['word_count_comment'].median() )
print ()
print("Среднее количество слов в  популярном комментарии (1 score) -", target_1['word_count_comment'].mean() )
print("Медианное количество слов в популярном комментарии (1 score) -", target_1['word_count_comment'].median() )
print()
print("Среднее количество слов в среднем по популярности комментарии (2 score) -", target_2['word_count_comment'].mean() )
print("Медианное количество слов в среднем по популярности комментарии (2 score) -", target_2['word_count_comment'].median() )
print()
print("Среднее количество слов в  непопулярном комментарии (3 score) -", target_3['word_count_comment'].mean() )
print("Медианное количество слов в непопулярном комментарии (3 score) -", target_3['word_count_comment'].median() )
print()
print("Среднее количество слов в самом непопулярном комментарии (4 score) -", target_4['word_count_comment'].mean() )
print("Медианное количество слов в самом непопулярном комментарии (4 score) -", target_4['word_count_comment'].median() )
print()
```

```
Среднее количество слов в самом популярном комментарии (0 score) - 141.18155197657393
Медианное количество слов в самом популярном комментарии (0 score) - 99.0

Среднее количество слов в  популярном комментарии (1 score) - 96.47291361639824
Медианное количество слов в популярном комментарии (1 score) - 67.0

Среднее количество слов в среднем по популярности комментарии (2 score) - 77.67884504068917
Медианное количество слов в среднем по популярности комментарии (2 score) - 52.0

Среднее количество слов в  непопулярном комментарии (3 score) - 66.69383817403839
Медианное количество слов в непопулярном комментарии (3 score) - 44.0

Среднее количество слов в самом непопулярном комментарии (4 score) - 57.975382205727313
Медианное количество слов в самом непопулярном комментарии (4 score) - 38.0
```

****** Average number of words in the most popular comment (0 score) - 141.18155197657393 Median number of words in the most popular comment (0 score) - 99.0

Average number of words in a popular comment (1 score) - 96.47291361639824 Median number of words in a popular comment (1 score) - 67.0

Average number of words on average by popularity comment (2 score) - 77.67884504068917 Median number of words on average by popularity comment (2 score) - 53.0

Average number of words in an unpopular comment (3 score) - 66.69383817403839 Median number of words in an unpopular comment (3 score) - 44.0

Average number of words in the most unpopular comment (4 score) - 57.975382205727313 Median number of words in the most unpopular comment (4 score) - 38.0

Conclusion for the first hypothesis: There is a significant difference between the number of words in popular and unpopular comments. The median number of words in the most popular comment for each post is almost twice that of the least popular.

Hypothesis 2: Popular comments (since they are more informative) use more complex, longer (multi-letter) words.

In [19]:
```
# calculate the average number of letters in words in comments
train_df['characters'] = train_df['comment_text'].str.len()
train_df['mean_of_characters_per_words'] = train_df['characters']/train_df['word_count_comment']
train_df.head(3)
```

Out[19]:

	post_text	comment_text	score	word_count_comment	characters	mean_of_characters_per_words
0	How many summer Y Combinator funders decided n...	Going back to school is not identical with giv...	0	186	998	5.365591
1	How many summer Y Combinator funders decided n...	There will invariably be those who don't see t...	1	76	414	5.447368
2	How many summer Y Combinator funders decided n...	For me school is a way to be connected to what...	2	91	488	5.362637

In [11]:
```
target_0 = train_df.loc[train_df['score'] == 0]
target_1 = train_df.loc[train_df['score'] == 1]
target_2 = train_df.loc[train_df['score'] == 2]
target_3 = train_df.loc[train_df['score'] == 3]
target_4 = train_df.loc[train_df['score'] == 4]
```

In [12]:
```
print("Среднее количество букв в используемых словах в самом популярном комментарии (0 score) -", round(target_0['mean_of_char
print("Медианное количество букв в используемых словах в самом популярном комментарии (0 score) -", round(target_0['mean_of_ch
print()
print("Среднее количество букв в используемых словах в  популярном комментарии (1 score) -", round(target_1['mean_of_characte
print("Медианное количество букв в используемых словах в популярном комментарии (1 score) -", round(target_1['mean_of_characte
print()
print("Среднее количество букв в используемых словах в среднем по популярности комментарии (2 score) -", round(target_2['mean_
print("Медианное количество букв в используемых словах в среднем по популярности комментарии (2 score) -", round(target_2['mean
print()
print("Среднее количество букв в используемых словах в  непопулярном комментарии (3 score) -", round(target_3['mean_of_charact
print("Медианное количество букв в используемых словах в непопулярном комментарии (3 score) -", round(target_3['mean_of_charac
print()
print("Среднее количество букв в используемых словах в самом непопулярном комментарии (4 score) -", round(target_4['mean_of_cha
print("Медианное количество букв в используемых словах в самом непопулярном комментарии (4 score) -", round(target_4['mean_of_
print()
```

```
Среднее количество букв в используемых словах в самом популярном комментарии (0 score) - 6.56
Медианное количество букв в используемых словах в самом популярном комментарии (0 score) - 5.9

Среднее количество букв в используемых словах в  популярном комментарии (1 score) - 6.56
Медианное количество букв в используемых словах в популярном комментарии (1 score) - 5.88

Среднее количество букв в используемых словах в среднем по популярности комментарии (2 score) - 6.56
Медианное количество букв в используемых словах в среднем по популярности комментарии (2 score) - 5.87

Среднее количество букв в используемых словах в  непопулярном комментарии (3 score) - 6.59
Медианное количество букв в используемых словах в непопулярном комментарии (3 score) - 5.85

Среднее количество букв в используемых словах в самом непопулярном комментарии (4 score) - 6.56
Медианное количество букв в используемых словах в самом непопулярном комментарии (4 score) - 5.82
```

****** Average number of letters in words used in the most popular comment (0 score) - 6.57 Median number of letters in words used in the most popular comment (0 score) - 5.9

Average number of letters in words used in a popular comment (1 score) - 6.56 Median number of letters in words used in a popular comment (1 score) - 5.88

Average number of letters in words used on average by popularity comment (2 score) - 6.58 Median number of letters in words used on average by popularity comment (2 score) - 5.87

Average number of letters in words used in an unpopular comment (3 score) - 6.59 Median number of letters in words used in an unpopular comment (3 score) - 5.85

Average number of letters in words used in the most unpopular comment (4 score) - 6.56 Median number of letters in words used in the most unpopular comment (4 score) - 5.82

Conclusion for the second hypothesis: we can find a consistent difference between the popularity of a comment and the length of the words that are used in it. Popular comments do use longer words compared to unpopular comments.

Hypothesis 3: Popular comments use more complex sentences. Compound sentences and enumerations are used more often. We will be able to see more commas in popular comments.

In [13]:
```
# number of commas in a comment and number of commas per number of words
train_df['commas'] = train_df['comment_text'].apply(lambda x : x.count(',') )
train_df['commas_per_words_count'] = train_df['word_count_comment']/train_df['commas']

train_df.head(3)
```

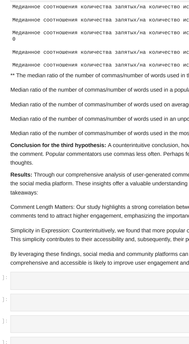
Out[13]:

	post_text	comment_text	score	word_count_comment	characters	mean_of_characters_per_words	commas	commas_per_words_count
0	How many summer Y Combinator funders decided n...	Going back to school is not identical with giv...	0	186	998	5.365591	7	26.571429
1	How many summer Y Combinator funders decided n...	There will invariably be those who don't see t...	1	76	414	5.447368	3	25.333333
2	How many summer Y Combinator funders decided n...	For me school is a way to be connected to what...	2	91	488	5.362637	0	inf

In [16]:
```
target_0 = train_df.loc[train_df['score'] == 0]
target_1 = train_df.loc[train_df['score'] == 1]
target_2 = train_df.loc[train_df['score'] == 2]
target_3 = train_df.loc[train_df['score'] == 3]
target_4 = train_df.loc[train_df['score'] == 4]
```

In [17]:
```
plt.clf()
sns.histplot(datartrain_df, x="commas_per_words_count", bins=1000, hue="score")
plt.xlim(0, 100)
plt.xlabel("Количество запятых на количество слов (соотношение) в комментарии")
plt.ylabel("Количество комментариев")
```

Out[17]:
```
Text(0, 0.5, 'Количество комментариев')
```



In [18]:
```
print("Медианное соотношение количества запятых/на количество используемых слов в самом популярном комментарии (0 score) -", r
print()
print("Медианное соотношение количества запятых/на количество используемых слов в популярном комментарии (1 score) -", round(t
print()
print("Медианное соотношение количества запятых/на количество используемых слов в среднем по популярности комментарии (2 score
print()
print("Медианное соотношение количества запятых/на количество используемых слов в непопулярном комментарии (3 score) -", round
print()
print("Медианное соотношение количества запятых/на количество используемых слов в самом непопулярном комментарии (4 score) -",
print()
```

```
Медианное соотношение количества запятых/на количество используемых слов в самом популярном комментарии (0 score) - 25.4

Медианное соотношение количества запятых/на количество используемых слов в популярном комментарии (1 score) - 27.33

Медианное соотношение количества запятых/на количество используемых слов в среднем по популярности комментарии (2 score) - 29.0

Медианное соотношение количества запятых/на количество используемых слов в непопулярном комментарии (3 score) - 31.0

Медианное соотношение количества запятых/на количество используемых слов в самом непопулярном комментарии (4 score) - 34.0
```

****** The median ratio of the number of commas/number of words used in the most popular comment (0 score) is 25.4

Median ratio of the number of commas/number of words used in a popular comment (1 score) - 27.33

Median ratio of the number of commas/number of words used on average by popularity of comments (2 score) - 29.0

Median ratio of the number of commas/number of words used in an unpopular comment (3 score) - 31.0

Median ratio of the number of commas/number of words used in the most unpopular comment (4 score) - 34.0

Conclusion for the third hypothesis: A counterintuitive conclusion, however, in more popular comments there are fewer commas per number of words in the comment. Popular commentators use commas less often. Perhaps fewer commas means a simpler and more understandable form of expressing thoughts.

Results: Through our comprehensive analysis of user-generated comments, we have identified key factors that significantly influence comment popularity on the social media platform. These insights offer a valuable understanding of the dynamics of user engagement and content optimization. Here are the main takeaways:

Comment Length Matters: Our study highlights a strong correlation between comment length and its likelihood of becoming popular. Longer, more elaborative comments tend to attract higher engagement, emphasizing the importance of well-expressed thoughts.

Simplicity in Expression: Counterintuitively, we found that more popular comments use fewer commas, indicating simpler and more straightforward language. This simplicity contributes to their accessibility and, subsequently, their popularity among a broader audience.

By leveraging these findings, social media and community platforms can enhance their content optimization strategies. Tailoring comments to be more comprehensive and accessible is likely to improve user engagement and foster effective communication within online communities.

In [ ]:
```

```

In [ ]:
```

```

In [ ]:
```

```

In [ ]:
```

```

In [ ]:
```

```

In [ ]:
```

```

In [ ]:
```

```