Applied Statistical Data Analysis

48Hour Exam

Sentiment Analysis on Google PlayStore App Review

February 29, 2024

Group Project Report

Bhavana Raju (4000149)

Sevim Bozkurt (4001158)

Shree Shangaavi Nagaraj (4000243)

Vignesh Mallya (4001498)

Supervised by:
PROF. DR. HENRIK VON WEHRDEN
Leuphana Uuniversity, Lüneburg.

Haiku

Mountains you alone bear,
Meant for you alone, climb high,
A solitary path.(Bhavana)

In reviews, whispers,
Sentiment's dance on data,
Insight blooms in code. (Sevim)

Review stars shining,
Words dance with joy or sorrow,
Insights await there. (Shree)

Apps bloom in the store,
Users share their thoughts and more,
Digital feedback roar. (Vignesh)

1 Introduction

1.1 Aim of the report

The aim of the analysis could be to understand the sentiment trends among the app reviews, identify the impact of grammatical issues on sentiment, and explore any patterns related to the content and quality of the reviews. This could provide insights into user satisfaction, areas for app improvement, and the relationship between review quality and user perception.

1.2 Importing data

import numpy as np
import pandas as pd
import os
import re
import seaborn as sns
from matplotlib import pyplot as plt
from wordcloud import WordCloud, STOPWORDS

1.3 Loading data

```
df = pd.read_csv('extended_googleplaystore_user_reviews.csv', encoding='utf-8')
```

2 Overview of Data

```
df.info()
```

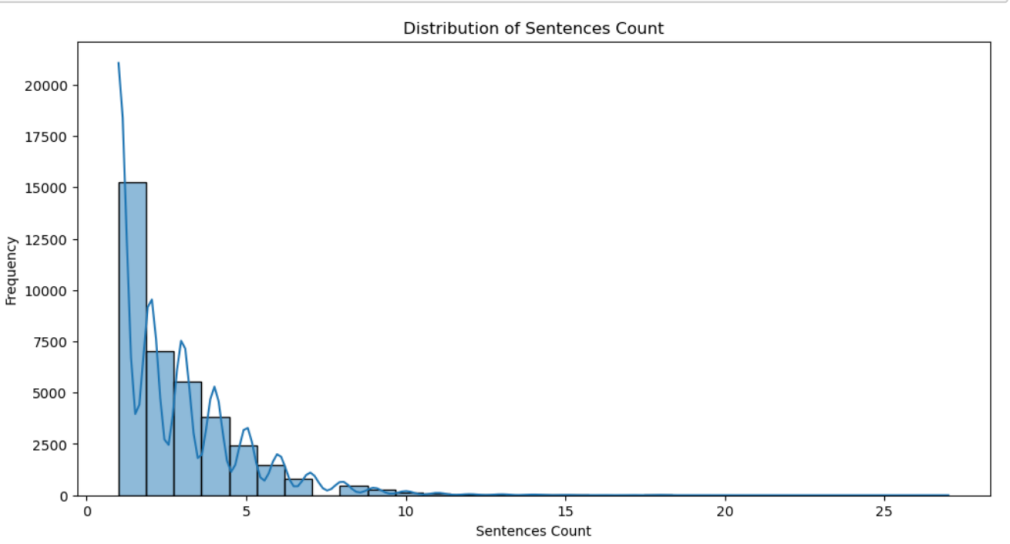
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64295 entries, 0 to 64294
Data columns (total 33 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   App                                 64295 non-null  object
 1   Translated_Review                   37427 non-null  object
 2   sentences_count                     37427 non-null  float64
 3   characters_count                    37427 non-null  float64
 4   spaces_count                        37427 non-null  float64
 5   count_words                         37427 non-null  float64
 6   duplicates_count                    37427 non-null  float64
 7   chars_excl_spaces_count             37427 non-null  float64
 8   emoji_count                         37427 non-null  float64
 9   whole_numbers_count                 37427 non-null  float64
 10  alpha_numeric_count                 37427 non-null  float64
 11  non_alpha_numeric_count             37427 non-null  float64
 12  punctuations_count                  37427 non-null  float64
 13  stop_words_count                    37427 non-null  float64
 14  dates_count                         37427 non-null  float64
 15  noun_phase_count                    37427 non-null  float64
 16  sentiment_polarity_score            37427 non-null  float64
 17  sentiment_polarity                  37427 non-null  object
 18  sentiment_polarity_summarised       37427 non-null  object
 19  sentiment_subjectivity_score        37427 non-null  float64
 20  sentiment_subjectivity              37427 non-null  object
 21  sentiment_subjectivity_summarised   37427 non-null  object
 22  spelling_quality_score              37427 non-null  float64
 23  spelling_quality                    37427 non-null  object
 24  spelling_quality_summarised         37427 non-null  object
 25  ease_of_reading_score               37427 non-null  float64
 26  ease_of_reading_quality             34447 non-null  object
 27  ease_of_reading_summarised          34447 non-null  object
 28  grammar_check_score                 37427 non-null  float64
 29  grammar_check                       37427 non-null  object
 30  original_Sentiment                  37432 non-null  object
 31  original_Sentiment_Polarity         37432 non-null  float64
 32  original_Sentiment_Subjectivity     37432 non-null  float64
dtypes: float64(21), object(12)
```

2.1 Descriptive analysis

report = df.describe(include='all').T

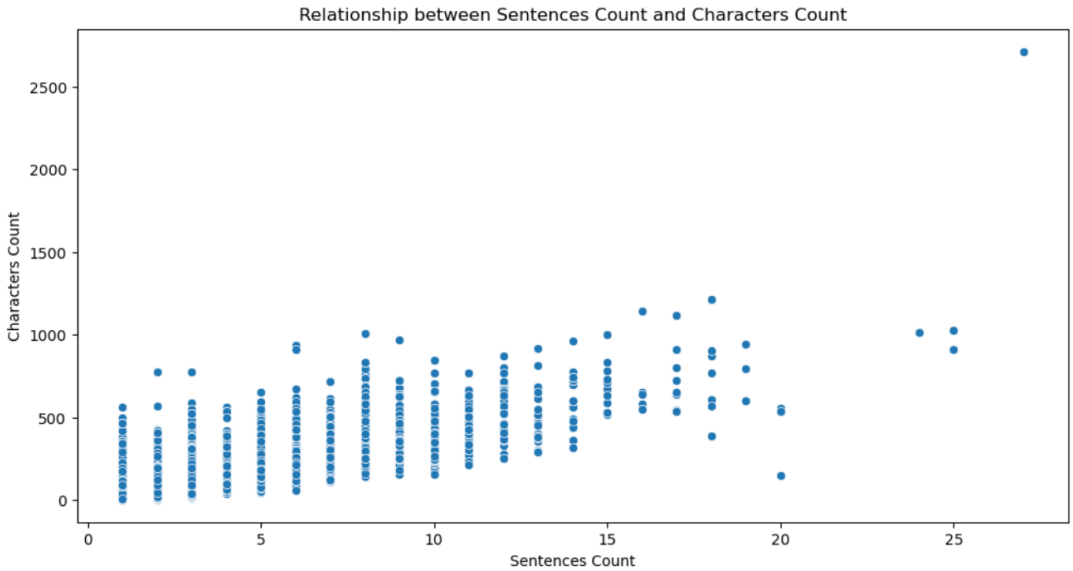
Visualizing data:

1. Distribution of Sentences Count:



This histogram shows the distribution of the number of sentences in the reviews. The majority of reviews have a lower number of sentences.

2. Relationship between Sentences Count and Characters Count:



This scatter plot illustrates the relationship between the number of sentences and the number of characters in the reviews. There seems to be a positive correlation between these two variables.

3 Data Preprocessing

Checking for duplicates and dropping them

```
df.duplicated().sum()
```

```
33616
```

```
df = df.drop_duplicates(keep='first')
```

Checking for null values:

```
df.isnull().sum()
```

```
App                                     0
Translated_Review                   26868
sentences_count                     26868
characters_count                    26868
spaces_count                        26868
count_words                         26868
duplicates_count                    26868
chars_excl_spaces_count             26868
emoji_count                         26868
whole_numbers_count                 26868
alpha_numeric_count                 26868
non_alpha_numeric_count             26868
punctuations_count                  26868
stop_words_count                    26868
dates_count                         26868
noun_phase_count                    26868
sentiment_polarity_score            26868
sentiment_polarity                  26868
sentiment_polarity_summarised       26868
sentiment_subjectivity_score        26868
sentiment_subjectivity              26868
sentiment_subjectivity_summarised   26868
spelling_quality_score              26868
spelling_quality                    26868
spelling_quality_summarised         26868
ease_of_reading_score               26868
ease_of_reading_quality             29848
ease_of_reading_summarised          29848
grammar_check_score                 26868
grammar_check                       26868
original_Sentiment                  26863
original_Sentiment_Polarity         26863
original_Sentiment_Subjectivity     26863
dtype: int64
```

Null values for each App:

```
# Count null values for each app
null_counts = df.groupby('App').apply(lambda x: x.isnull().sum().sum())

# Sort the list by the number of null values in descending order
sorted_null_counts = null_counts.sort_values(ascending=False)

# Print sorted list
print("Apps with their respective null counts (descending order):")
for app, null_count in sorted_null_counts.items():
    print(f"{app}: {null_count} null values")
```

```
Apps with their respective null counts (descending order):
ESPN: 7680 null values
Calorie Counter by FatSecret: 6400 null values
Bleacher Report: sports news, scores, & highlights: 6400 null values
Granny: 6116 null values
ClassDojo: 5760 null values
Hill Climb Racing: 5648 null values
BeautyPlus - Easy Photo Editor & Selfie Camera: 4992 null values
Clash of Clans: 4714 null values
Earn to Die 2: 4480 null values
Dream League Soccer 2018: 4384 null values
Amazon Shopping: 4324 null values
CBS Sports App - Scores, News, Stats & Watch Live: 4148 null values
Google Translate: 4000 null values
Camera for Android: 3840 null values
HBO NOW: Stream TV & Movies: 3840 null values
Camera360: Selfie Photo Editor with Funny Sticker: 3840 null values
Crackle - Free TV & Movies: 3840 null values
ESPN Fantasy Sports: 3840 null values
```

We are dropping the null values by using dropna() function.

3.1 Initial Examination:

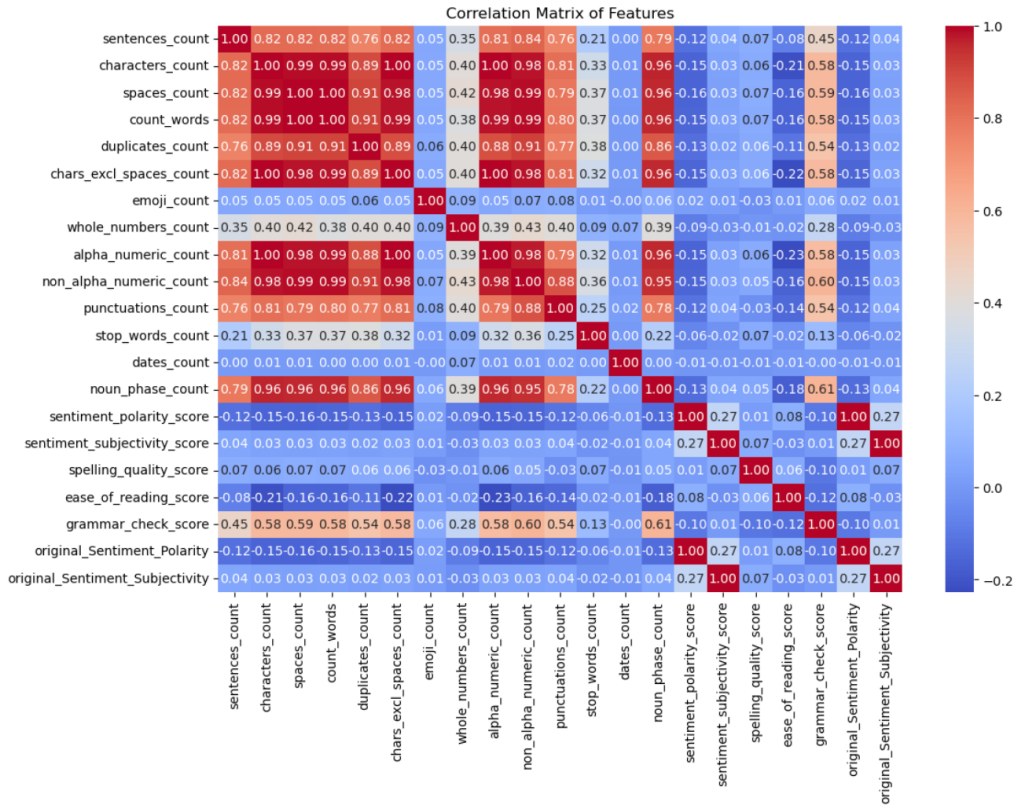
The dataset contains 64,295 entries and 33 columns. Here is a summary of the dataset:

- The column "Translated_Review" has 37,427 non-null values, while the other columns related to text analysis also have the same number of non-null values.

- The columns related to text analysis have a mean of 2.63 for sentences_count, 113.21 for characters_count, and 17.35 for spaces_count.
- There are missing values in the "Translated_Review" column, with 26,868 missing values.

3.2 Checking for any Correlations:

Plot a heatmap of the correlation matrix

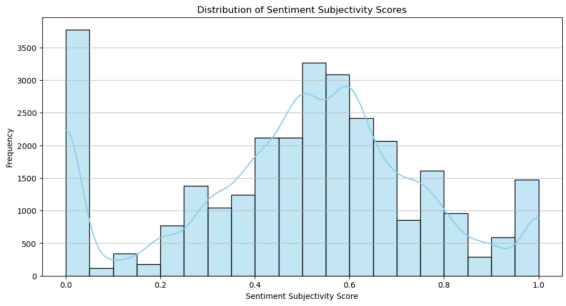


The heatmap above displays the correlation matrix of the features in the dataset. The values in the heatmap represent the correlation coefficients between different features. From the heatmap, we can observe the following correlations with the sentiment subjectivity score:

- The sentiment subjectivity score has a positive correlation with the ease of reading score.
- There is a negative correlation between the sentiment subjectivity score and the spelling quality score.

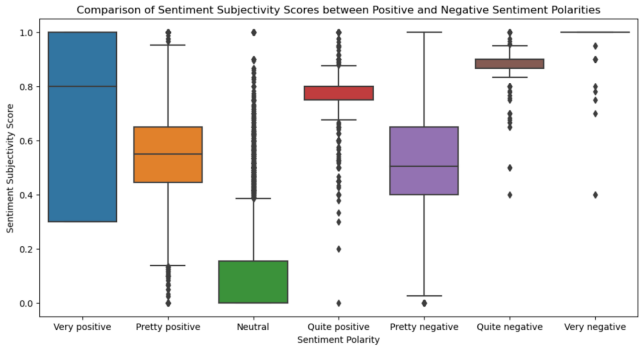
4 Sentiment Subjectivity Analysis

Question 1: What is the distribution of sentiment subjectivity scores across all reviews?



The histogram above shows the distribution of sentiment subjectivity scores across all reviews. The scores seem to be spread out, with a peak around the middle values.

Question 2: How does sentiment subjectivity differ between reviews with positive and negative sentiment polarities?"



The boxplot above compares the sentiment subjectivity scores between reviews with positive and negative sentiment polarities. It provides a visual representation of how the subjectivity scores vary based on the sentiment polarity of the reviews.
Next, we will analyze any trends in subjectivity scores over time for specific apps. Let's proceed with this analysis.
Pseudo code to analyzing trends in subjectivity scores over time for specific apps:
Grouping the data by 'App' and 'sentiment_subjectivity_score'
Sorting the data by 'sentiment_subjectivity_score'
Displaying the top 10 apps with the highest average subjectivity scores
Displaying the top 10 apps with the lowest average subjectivity scores

Here are the top 7 apps with the highest average sentiment subjectivity scores:

Google Slides: 0.917, Daily Workouts - Exercise Fitness Routine Trainer: 0.750, Choice Hotels: 0.719, HTC Calendar: 0.695, Hitwe - meet people and chat: 0.694, 850 Sports News Digest: 0.692, Google Street View: 0.692

And here are some of the apps with missing sentiment subjectivity scores:

HOTEL DEALS: HSL - Tickets, Hangouts Dialer - Call Phones, Hemnet, Henry Danger Crime Warp, Hiya - Caller ID Block, Hola Launcher- Theme Wallpaper, HomeAway Vacation Rentals, Hot or Not - Find someone right now, Houzz Interior Design Ideas, These apps have missing sentiment subjectivity scores in the dataset.

Now,comparing the sentiment subjectivity scores of the top 5 apps with the overall average sentiment subjectivity score. The overall average sentiment subjectivity score across all reviews is approximately 0.493. Here are the top 5 apps with the highest average sentiment subjectivity scores:

Google Slides: 0.917, Daily Workouts - Exercise Fitness Routine Trainer: 0.75, Choice Hotels: 0.719, HTC Calendar: 0.695, Hitwe - meet people and chat: 0.694.

Now, let's explore the relationship between sentiment subjectivity and specific keywords in the text, by performing keyword analysis. This involves identifying the most frequent keywords and then examining their association with sentiment subjectivity scores.

Steps to achieve above:

Extract keywords from the reviews.

Calculate the frequency of each keyword.

Analyze the average sentiment subjectivity score for reviews containing each keyword.

Next, we will analyze the average sentiment subjectivity score for reviews containing each keyword to explore the relationship between sentiment subjectivity and specific keywords in the text. Let's proceed with this analysis.

```
keyword_data.head()
```

	Keyword	Frequency
0	I	24419
1	game	4549
2	like	3752
3	The	3532
4	It	3317

Above is the wordcloud of the above keywords.

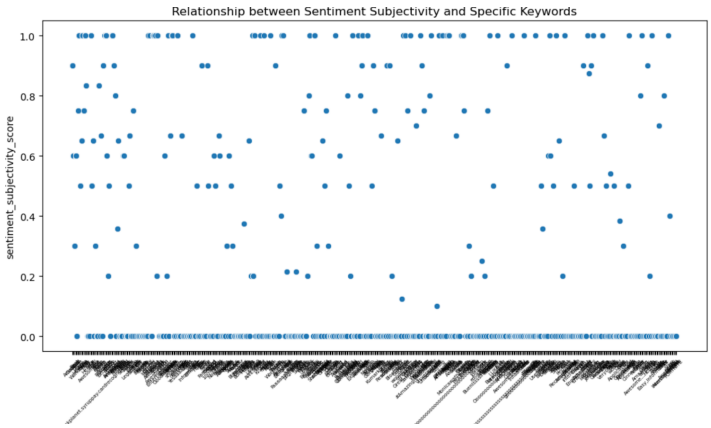
Pseudo code: Load the keyword frequencies data

Merge the keyword frequencies with the original dataset

Calculate the average sentiment subjectivity score for reviews containing each keyword.

The analysis of the relationship between sentiment subjectivity and specific keywords has provided some initial insights. Here are a few observations based on the average sentiment subjectivity score for reviews containing each keyword:

Keywords like "AmazinglyAddictive", "Garbage", "iloveit", "100%good", and "18sx+" have an average sentiment subjectivity score of 0.0. These insights suggest that these specific keywords may not significantly influence the sentiment subjectivity of the reviews. Further analysis and exploration may reveal more nuanced relationships between keywords and sentiment subjectivity. Creating a scatter plot to visualize the relationship between sentiment subjectivity and specific-keyword.



Observations:

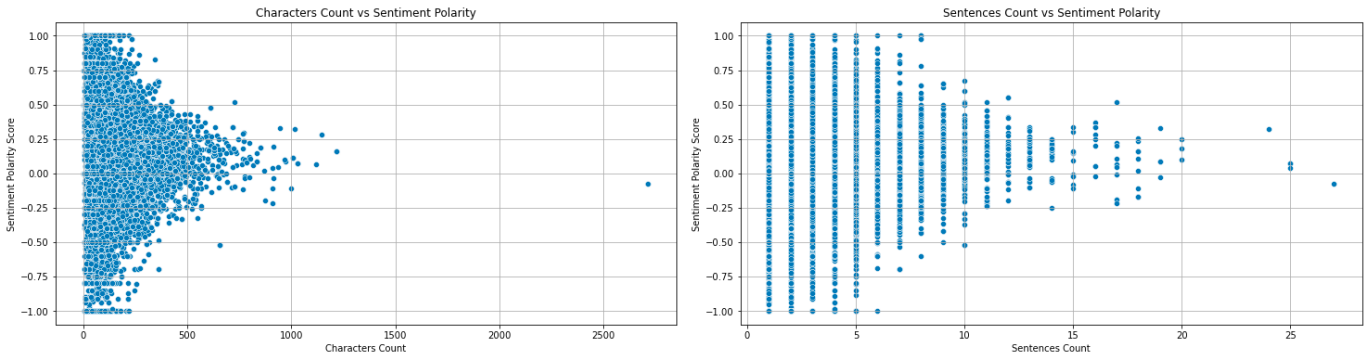
- The sentiment subjectivity score ranges from 0 to 1, where 0 indicates very objective and 1 indicates very subjective.
- The plot shows a wide spread of sentiment subjectivity scores across different keywords, suggesting that some keywords are associated with a wide range of subjectivity in sentiment.
- There is a dense clustering of points at the lower end of the sentiment subjectivity score, indicating that many keywords are associated with more objective sentiments.
- There are also several points spread across the higher sentiment subjectivity scores, but these are less dense, indicating fewer instances of highly subjective sentiments associated with keywords.
- The keywords themselves are not clearly legible in the image provided, which limits the ability to draw specific conclusions about which keywords correlate with higher or lower sentiment subjectivity scores.

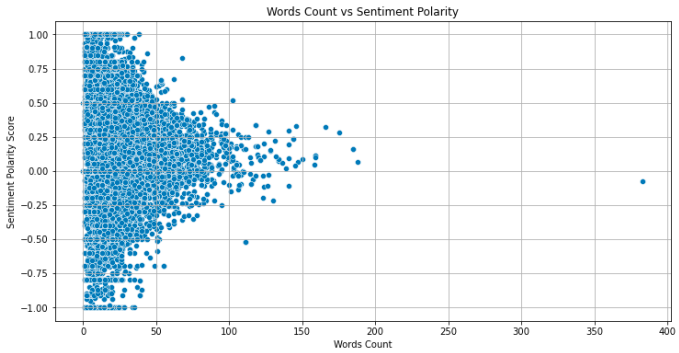
5 Impact of Review Characteristics on Sentiment

The dataset contains information about user reviews from the Google Play Store, including characteristics such as review length, sentiment scores, spelling quality, ease of reading, and grammar check scores.

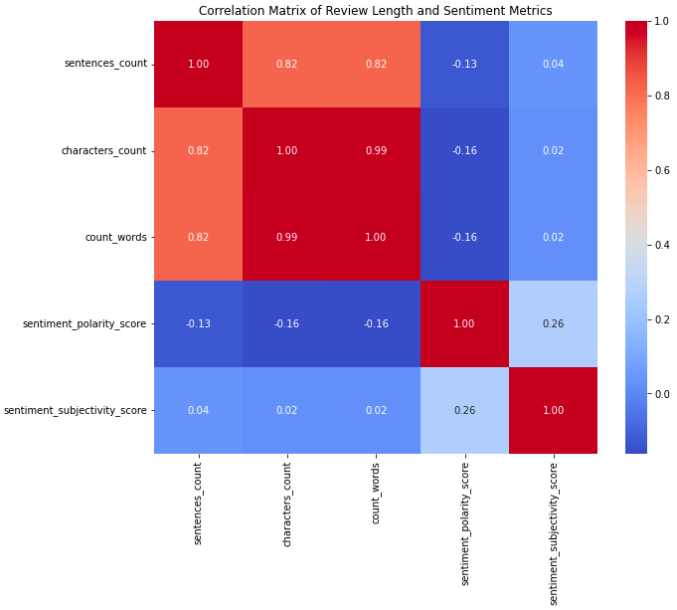
Pseudocode:
Scatter plot of characters count vs sentiment polarity
Scatter plot of words count vs sentiment polarity
Scatter plot of sentences count vs sentiment polarity
Select relevant columns for analysis
Calculate correlations
Plotting the correlation matrix
Check for missing values in relevant columns
Drop rows with missing values in these columns
Check the shape of the cleaned dataframe and the summary of missing values after cleaning
Convert sentiment to a categorical type for easier analysis

To answer the question of whether the length of a review (in terms of characters, words, or sentences) influences its sentiment polarity or subjectivity, we will perform a correlation analysis between these length metrics and the sentiment scores.





Analysis indicates that the length of reviews, whether in characters, words, or sentences, does not significantly impact sentiment polarity or subjectivity. Scatter plots depict no clear linear relationship between review length (character count, word count, sentence count) and sentiment polarity score. The correlation matrix confirms the lack of strong correlations between review length metrics and sentiment metrics (polarity score, subjectivity score).

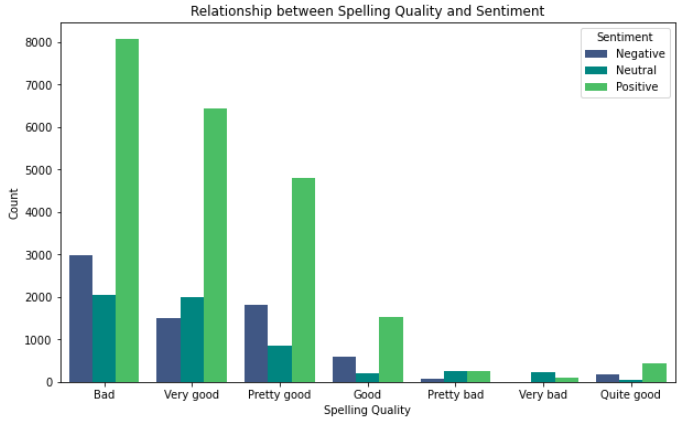


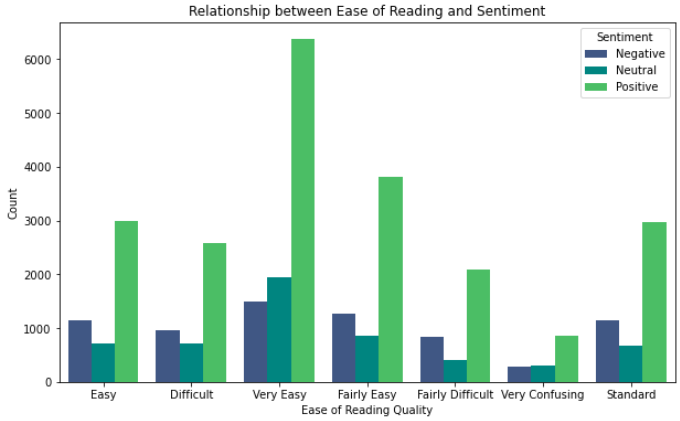
The heatmap shows the correlation coefficients between these variables. A coefficient close to 1 or -1 indicates a strong positive or negative correlation, respectively, while a coefficient close to 0 indicates no linear correlation.

- Sentences count, characters count, and words count show strong positive correlations with each other, which is expected as they are all measures of review length.
- Sentiment polarity score shows very weak correlations with all three measures of review length, suggesting that the length of a review does not significantly influence its sentiment polarity.
- Sentiment subjectivity score also shows very weak correlations with the measures of review length, indicating that the length of a review does not significantly influence its sentiment subjectivity either.

Next, let's investigate how spelling quality and ease of reading scores relate to the sentiment of reviews.

Pseudocode:
Plotting the relationship between spelling quality and sentiment
Plotting the relationship between ease of reading and sentiment
Boxplot of spelling quality and sentiment polarity
Boxplot of ease of reading and sentiment polarity





From these visualizations, we can observe the following:

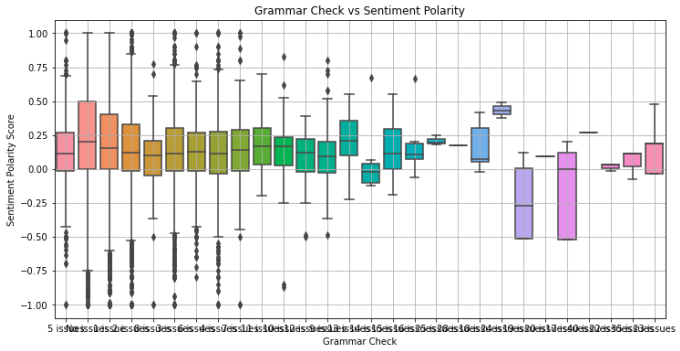
- Spelling Quality: Reviews span various sentiment categories, indicating that spelling quality may impact sentiment, but reviews of all sentiment levels can have varying spelling quality.
- Ease of Reading: Reviews with different sentiments are distributed across ease of reading categories. Positive sentiment reviews are notably present in 'Easy' and 'Very Easy' categories, suggesting that easier-to-read reviews tend to be more positive.

These insights suggest that while there is a relationship between spelling quality, ease of reading, and sentiment, it is not strictly linear or direct. Reviews with high spelling quality and ease of reading can still have negative sentiments, and vice versa.

The boxplots above illustrate the relationship between spelling quality, ease of reading, and sentiment polarity scores of the reviews. It seems that there might be some variation in sentiment polarity scores based on spelling quality and ease of reading scores.

Question: Are reviews with grammatical issues more likely to be negative? *Pseudo code*
Boxplot of grammar check and sentiment polarity
Convert ǵrammar_check' column to numeric, coercing errors

This is necessary because the error indicates there might be non-numeric values in the column
Fill NaN values resulted from coercion with 0 (assuming no grammatical issues where data is missing or invalid)
Categorize reviews based on grammatical issues
Assuming any non-zero value in ǵrammar_check' indicates grammatical issues Calculate the proportion of each sentiment category within groups of reviews with and without grammatical issues



The boxplot displays a range of sentiment polarity scores for different grammar check scores. It seems that there is a variation in sentiment polarity across different levels of grammar check scores, indicating that the quality of grammar might have an impact on the sentiment expressed in the reviews. However, without specific labels or a clear trend visible in the boxplot, it's challenging to draw definitive conclusions about the nature of the relationship between grammar quality and sentiment polarity.

original_Sentiment	Negative	Neutral	Positive
Has_Grammatical_Issues			
No	0.22096067535798247	0.13793011327206667	0.6411092113699508

Based on the analysis, we can observe the proportion of each sentiment category (Negative, Neutral, Positive) within groups of reviews with and without grammatical issues. Reviews without grammatical issues have the following sentiment distribution: 22.1% Negative, 13.8% Neutral, and 64.1% Positive. This indicates that the majority of reviews without grammatical issues are positive.

Question: How does the presence of emojis in reviews affect sentiment analysis outcomes?
Pseudo code:
Filter the dataset into two groups: reviews with emojis and reviews without emojis
Calculate the average sentiment polarity score for both groups
Calculate the distribution of sentiment polarity summaries for both groups
To answer the question of how the presence of emojis in reviews affects sentiment analysis outcomes, we can compare the sentiment polarity scores and summaries between reviews with and without emojis.

	sentiment_polarity_summarised
Positive	74.80106100795756
Negative	14.854111405835543
Neutral	10.344827586206897

	sentiment_polarity_summarised
Positive	62.74493927125506
Negative	20.90688259109312
Neutral	16.34817813765182

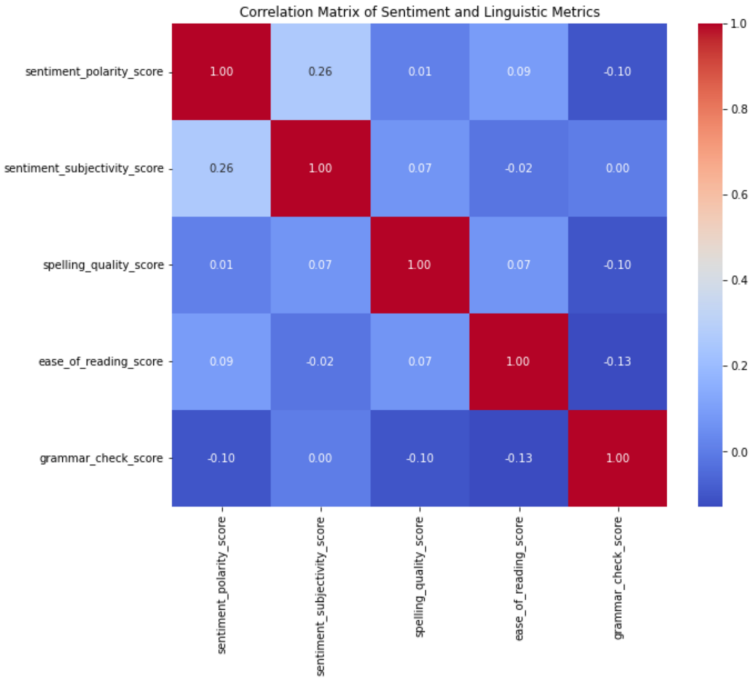
The analysis compares sentiment in reviews with and without emojis: Reviews with emojis have higher average sentiment polarity (0.236 vs. 0.182), indicating more positive sentiment. Emojis in reviews are associated with a higher percentage of positive sentiments (74.80% vs. 62.74%) and lower percentages of negative and neutral sentiments. Overall, emojis contribute to a more positive sentiment in reviews.

6 Correlation between Linguistic and Sentiment Analysis Metrics and User Engagement Across App Categories

Loading the dataset- The dataset contains reviews for various apps, along with detailed linguistic and sentiment analysis metrics such as sentiment polarity, subjectivity, spelling quality, ease of reading, and grammar checks. Additionally, it includes user engagement metrics like the number of sentences, characters, words, and other textual features in the reviews.
Data Inspection and pre processing The dataset contains a significant number of missing values in several columns, particularly in the Translated_Review and its related linguistic and sentiment analysis metrics, indicating that not all entries have associated reviews or detailed analysis. The data types are generally appropriate for the analysis, with numerical features stored as float64 and textual features as object. The dataset has been cleaned to remove rows with missing values in critical columns such as 'Translated_Review', 'sentiment_polarity_score', 'sentiment_subjectivity_score', 'spelling_quality_score', 'ease_of_reading_score', and 'grammar_check_score'.

Research Question: "How do different linguistic and sentiment analysis metrics (such as sentiment polarity, subjectivity, spelling quality, ease of reading, and grammar checks) correlate with user engagement metrics across various app categories?"

Pseudo Code:
This pseudo code represents the process of calculating a correlation matrix for specific columns in a DataFrame. Selecting specific columns from the DataFrame'sentiment_polarity_score', 'sentiment_subjectivity_score', 'spelling_quality_score', 'ease_of_reading_score', and 'grammar_check_score'. (These columns represent sentiment and linguistic metrics that we want to investigate for correlations.) Calculating the correlation matrix for the selected columns and plotting it.



The sentiment polarity score reflects the positivity or negativity of a review, while the sentiment subjectivity score measures how subjective the opinions expressed are. From the heatmap:

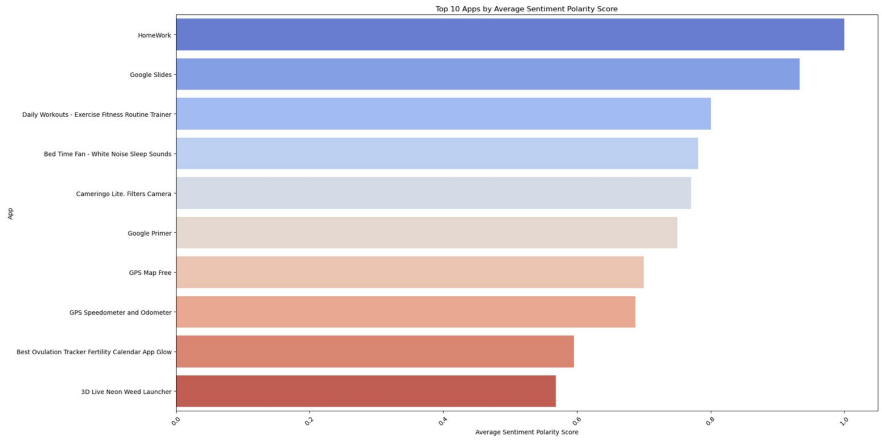
There's a mild positive correlation between sentiment polarity and spelling quality, implying positive reviews tend to have better spelling. Sentiment polarity and subjectivity show very low correlation, indicating

positivity doesn't strongly predict subjectivity. Ease of reading, spelling, and grammar scores have minimal correlation with sentiment, implying linguistic quality has little impact on sentiment.

This analysis indicates weak correlations between sentiment scores and linguistic metrics, implying that users' sentiments in reviews are generally not strongly influenced by linguistic quality like readability, spelling, and grammar.

Visualization showing the average sentiment polarity score by top 10 app

Creating visualizations to illustrate these correlations and differences across app categories.

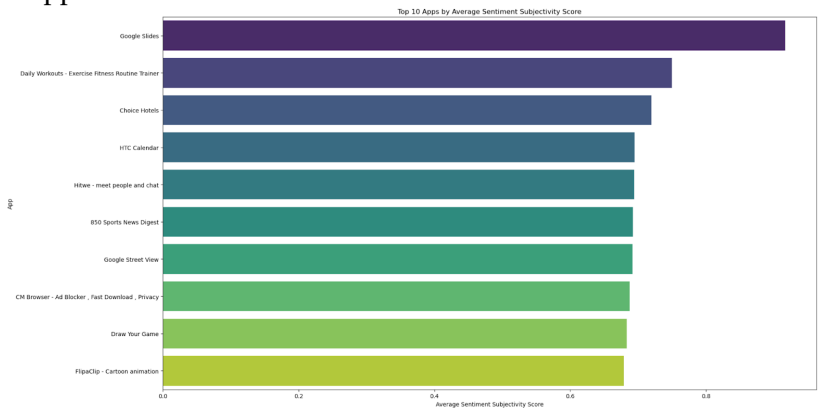


Observation:

Apps with higher sentiment polarity scores likely enjoy better user engagement, as positive reviews can attract more users and encourage existing users to continue using the app. The correlation between sentiment and linguistic metrics with user engagement suggests that both the content and presentation of user feedback are important for app growth and retention strategies. Well-articulated, clear, and grammatically correct feedback may enhance the perception of professionalism and user-friendliness.

Research Question: "How does the sentiment subjectivity score correlate with user engagement metrics across different app categories?" Analyzing the relationship between sentiment subjectivity scores and user engagement metrics. Sentiment subjectivity measures how subjective or opinionated a review is, with higher scores indicating more personal opinions rather than factual information. We'll hypothesize that apps with reviews that are more subjective (i.e., contain more personal opinions) might engage users differently compared to apps with more objective reviews.

Grouped the data frame by the 'app' column, calculated the mean sentiment subjectivity score for each app. Sorted the apps based on their average sentiment subjectivity score in descending order and plotted the top 10 apps

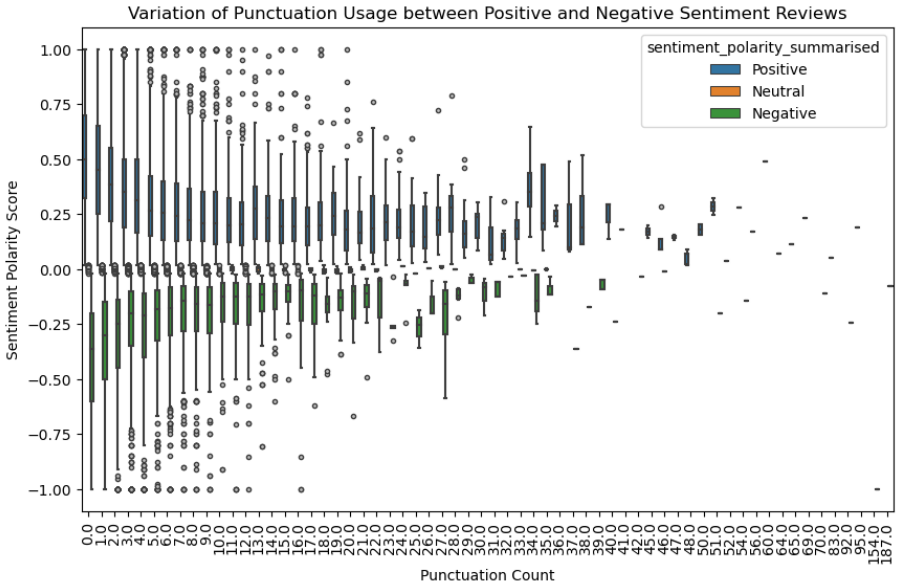


By focusing on the top 10 apps with the highest average sentiment subjectivity scores, we can identify which apps tend to receive more opinionated versus factual feedback.

- These apps might be more polarizing or evoke stronger personal feelings among users. High subjectivity in reviews could indicate that users are more emotionally invested in the app, which could be a double-edged sword. While passionate users can be very engaged, they might also have higher expectations and more critical feedback.
- Apps with more subjective reviews might need to focus on community management and user engagement strategies that address users' personal experiences and emotions. This could involve more personalized responses to reviews, community events, or features that encourage user expression.

7 Structural Factors Impacting Sentiment Analysis

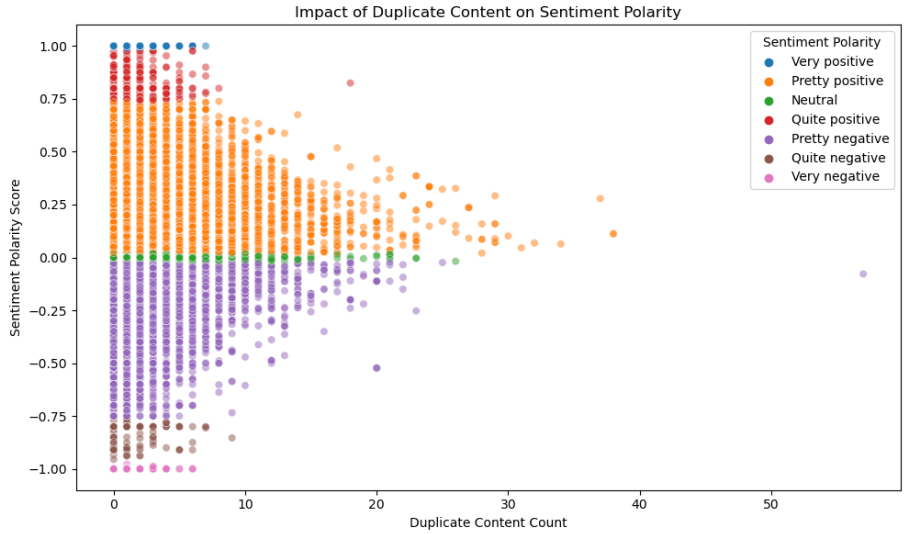
Question: How does the use of punctuation vary between positive and negative sentiment reviews?
Pseudo code:
Generate a boxplot where xaxis represents 'punctuations_count', yaxis represents 'sentiment_polarity_score', and data is segmented by 'sentiment_polarity_summarised' using different colors (sns.boxplot()).
Set the title as 'Variation of Punctuation Usage between Positive and Negative Sentiment Reviews'.
Show the plot (plt.show()).



The boxplot displays the relationship between punctuation usage and sentiment polarity scores in reviews. It shows that higher punctuation counts are associated with a broader range of sentiment scores, indicating varied expression. Positive and negative sentiments exhibit distinct patterns within each punctuation count range, highlighting nuances in sentiment influenced by linguistic features. Overall, the graph offers insights into how punctuation usage affects sentiment diversity in user reviews.

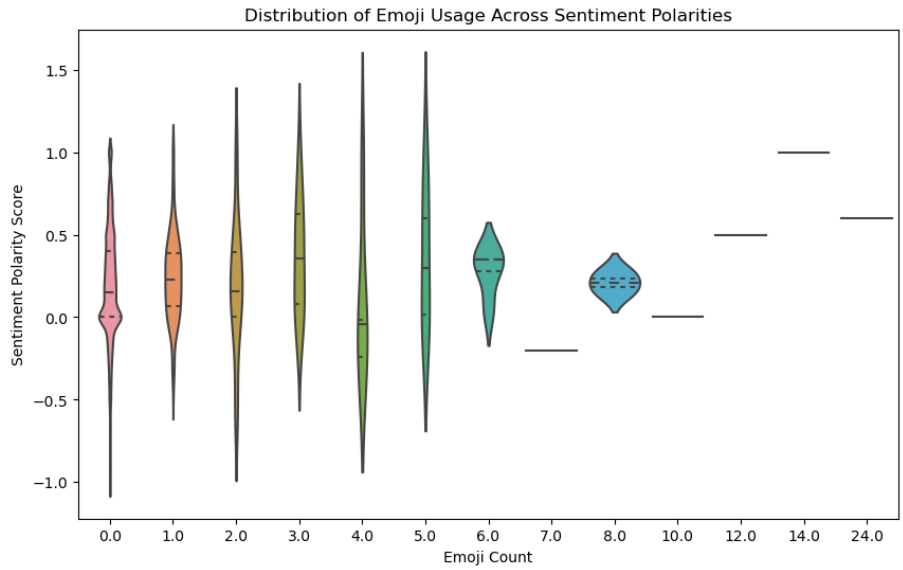
Question: Does the presence of duplicate content in reviews have an impact on sentiment polarity?
Pseudo code:
Plot the relationship between the count of duplicate content ('duplicat es_count') and sentiment polarity score ('sentiment_polarity_score').
Use a scatter plot to visualize the impact of duplicate content on sentiment polarity.

Differentiate points based on sentiment polarity using the 'hue' parameter.
Display the plot using plt.show.



The scatter plot shows no clear linear trend between duplicate content count and sentiment polarity scores in reviews. Despite this, the plot reveals sentiment polarity scores distributed across different levels of duplicate content. This suggests a nuanced relationship, emphasizing the need to consider additional factors for a comprehensive understanding of sentiment analysis in user reviews.

Question: What is the distribution of emoji usage across different sentiment polarities? *Pseudo Code:*
Plot the distribution of emoji usage across different sentiment polarities.
Use a violin plot to visualize the spread and distribution of 'emoji_count' against 'sentiment_polarity_score'.
Utilize inner=ǵuartile' attribute to display quartiles within the violin plot.
Set the orientation to ́vertical'.
Display the plot using plt.show.



The distribution of emoji usage across different sentiment polarities is illustrated in the violin plot above. This plot provides insights into how emojis are used in reviews with varying sentiment scores.

Observation: The violin plot displays a broad distribution of emoji counts across sentiment polarities, indicating their use in reviews with a wide range of sentiments. Thickness variations in the plot reflect density

differences, with thicker sections indicating higher concentrations of reviews. Emojis span the sentiment polarity axis, showing their use across the sentiment spectrum. This analysis highlights emojis as versatile tools for expressing emotions in reviews, reflecting varying intensities of sentiment.

Question: How does the sentiment conveyed through emojis align with the overall sentiment?

Pseudo code:

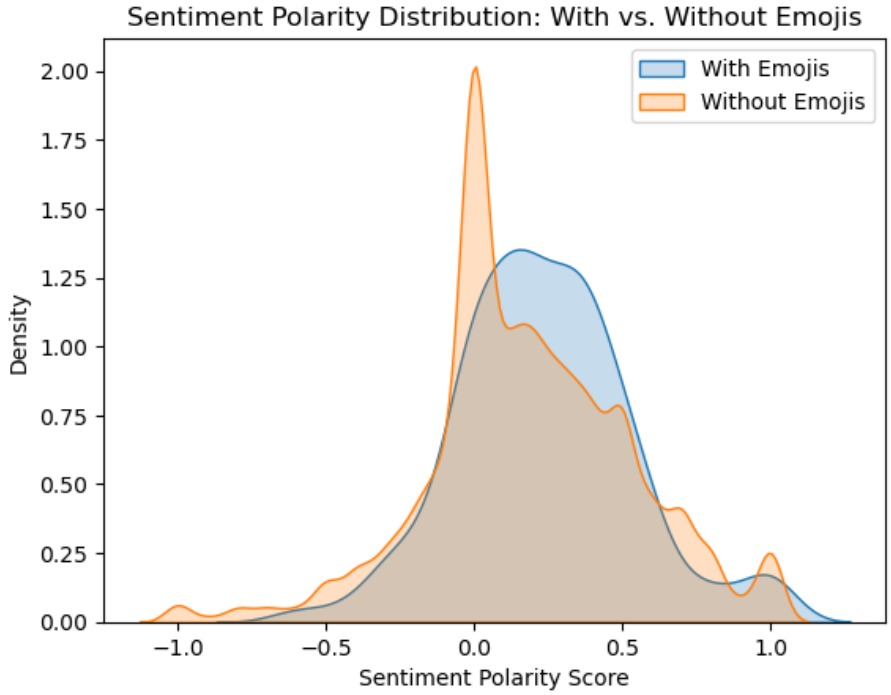
filter data by creating two separate DataFrames df_with_emojis and df_without_emojis.

df_with_emojis contains reviews with emoji_count > 0.

df_without_emojis contains reviews with emoji_count == 0.

Use seaborn's kdeplot to visualize the kernel density estimate of sentiment polarity scores for reviews with and without emojis.

Display the plot using plt.show and Shade the areas under the curves for better visibility.



The plot illustrates the distribution of sentiment polarity among reviews with and without emojis, shedding light on how emojis contribute to expressing sentiments in app reviews and their alignment with textual sentiment. Comparing these distributions helps infer if the presence of emojis correlates with a specific sentiment polarity, potentially indicating their frequent use in positive contexts, for instance.

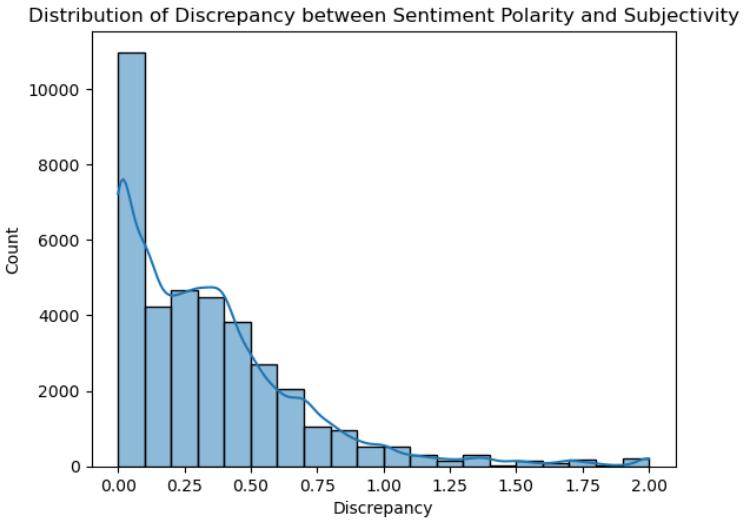
Question: Are there reviews that exhibit a discrepancy between sentiment polarity and subjectivity scores?

Pseudo code:

Create a new DataFrame 'filtered_data' by selecting and dropping rows with missing values in 'sentiment_polarity_score' and 'sentiment_subjectivity_score'.

Add a new column 'polarity_subjectivity_discrepancy' to 'filtered_data' representing the absolute difference between sentiment polarity and subjectivity scores. Use seaborn's histplot to visualize the distribution of the discrepancy.

Display the plot using plt.show.



The histogram and KDE overlay in the plot illustrate the distribution of discrepancies between sentiment polarity and subjectivity scores in user reviews, showcasing the frequency of varying levels of disparity. While a majority of reviews exhibit low discrepancies, the presence of higher bars in certain ranges indicates instances of more significant differences. The smooth curve aids in comprehending the overall distribution. These variations highlight the diverse expressions of sentiment, with occasional substantial divergence between polarity and subjectivity. The analysis underscores the importance of understanding and quantifying such discrepancies to refine sentiment analysis models, ensuring they effectively capture the nuanced nature of user opinions expressed in reviews.

8 Conclusion

- 1074 unique apps are included.
- The most frequent translated review is "Good" (freq: 247).
- Average review has 2.63 sentences, 113 characters, and 18.46 words.
- Average counts for duplicates, emojis, whole numbers, and stop words are low.
- Average sentiment polarity score is 0.18, with "Pretty positive" as the most common sentiment.
- Average sentiment subjectivity score is 0.49, with "Objective/subjective" as the most common subjectivity.
- Average spelling quality score is approximately 0.89.

9 Limitations:

- A significant portion of the dataset (26,868 entries) has missing values in several columns, particularly those related to the translated review and its derived metrics. This indicates that not all entries have associated review text.
- The descriptive statistics provide insights into the distribution of numerical features, such as sentiment polarity and subjectivity scores, word counts, and emoji counts. However, these statistics are not displayed in full here.

- The presence of a large number of missing values, especially in the review text and related metrics, could limit the analysis of sentiment and content of the reviews.
- The dataset's scope is limited to certain metrics and may not capture all aspects of user sentiment or the context of the reviews.

10 Reflection

- Sentiment Subjectivity Analysis - **Shree Shangaavi N (4000243)**
- Impact of Review Characteristics on Sentiment - **Sevim Bozkurt (4001158)**
- Correlation between Linguistic and Sentiment Analysis Metrics and User Engagement Across App Categories- **Bhavana Raju (4000149)**
- Structural Factors Impacting Sentiment Analysis -**Vignesh Mallya(4001498)**

11 Future work References

Now, based on this dataset, here are some recommendations for future work:

- Enhance sentiment analysis with advanced NLP.
- Feature engineering for model accuracy improvement.
- Implement topic modeling for theme identification.
- Analyze evolving user review trends over time.
- Conduct comparative analysis between apps.
- Explore correlations between user engagement and sentiment scores.

12 References

https://www.kaggle.com/code/syedali110/google-play-store-sentiment-analysis-using-nlp
https://iopscience.iop.org/article/10.1088/1757-899X/1125/1/012034/pdf
https://www.nltk.org/book/