CMPT 353Project
What Specific Linguistic Characteristics Distinguish Posts in the PCGaming subreddit from those in the MobileGaming subreddit?

Created By:
Yanjun Qian
Yinxuan Zhang

Introduction

The gaming community on Reddit is diverse, with subreddits dedicated to different aspects of gaming. Among these, the PCGaming and MobileGaming subreddits represent two distinct segments of the gaming population. While both communities focus on gaming, the platforms they discuss (PC vs. mobile) might influence their communication styles. This project aims to identify and analyze the specific linguistic characteristics that distinguish posts in the PCGaming subreddit from those in the MobileGaming subreddit. Specifically, we will analyze the frequency of words, sentiment, and readability of the submissions. By collecting and analyzing data from both subreddits, we aim to uncover significant linguistic patterns and differences, providing insights into how platform-specific factors affect online discussions and communication styles within these communities.

Problem Definition

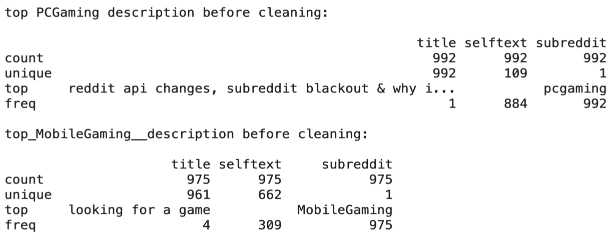
Our project aimed to refine the initial idea of comparing linguistic characteristics by focusing on three key aspects: word frequency, sentiment, and readability. We hypothesized that the different platforms (PC vs. mobile) might lead to distinct communication styles, reflected in the language used in each subreddit.

Data Collection

The primary data sources for this project are the PCGaming and MobileGaming subreddits on Reddit. Using the PRAW (Python Reddit API Wrapper) to access Reddit's API, we extracted 1,000 posts from each subreddit using the "top" filter with a time range of one year, ensuring we captured posts with the highest engagement from the previous year. For each post, we collected attributes such as title, selftext, score, ID, subreddit name, URL, number of comments, creation date and time, author's username, post type (self-post or link), NSFW flag, spoiler flag, upvote ratio, and vote counts. This dataset provides a detailed view of posts and their engagement within the two subreddits, enabling a thorough linguistic analysis.

Data Cleaning

Even though we initially filtered 1,000 posts in the data collection phase, the dataset consisted of 992 posts from the PCGaming subreddit and 976 posts from the MobileGaming subreddit (Figure 1). To ensure data quality, we removed posts with duplicate titles. Additionally, the PCGaming subreddit had many empty selftext fields. Therefore, we restricted our dataset to the top 900 submissions from each subreddit with unique titles. Due to the lack of unique selftext entries, especially in the PCGaming subreddit, we focused solely on the title data for further analysis. This cleaning process resulted in a robust dataset for reliable linguistic analysis (Figure 2).



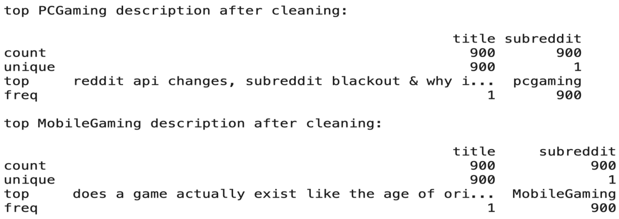
```
top PCGaming description before cleaning:

                                                    title selftext subreddit
count                                                 992      992       992
unique                                                992      109         1
top     reddit api changes, subreddit blackout & why i...           pcgaming
freq                                                    1      884       992

top_MobileGaming__description before cleaning:

                     title selftext    subreddit
count                  975      975          975
unique                 961      662            1
top     looking for a game          MobileGaming
freq                     4      309          975
```

Figure 1



```
top PCGaming description after cleaning:

                                                    title subreddit
count                                                 900       900
unique                                                900         1
top     reddit api changes, subreddit blackout & why i...  pcgaming
freq                                                    1       900

top MobileGaming description after cleaning:

                                                    title     subreddit
count                                                 900           900
unique                                                900             1
top     does a game actually exist like the age of ori...  MobileGaming
freq                                                    1           900
```

Figure 2

Data Processing

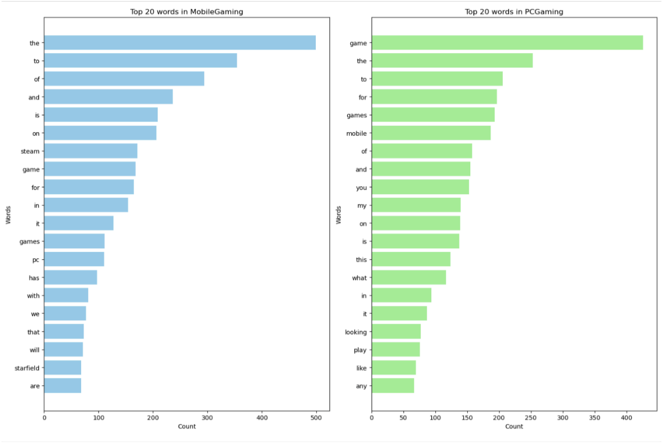


Figure 3

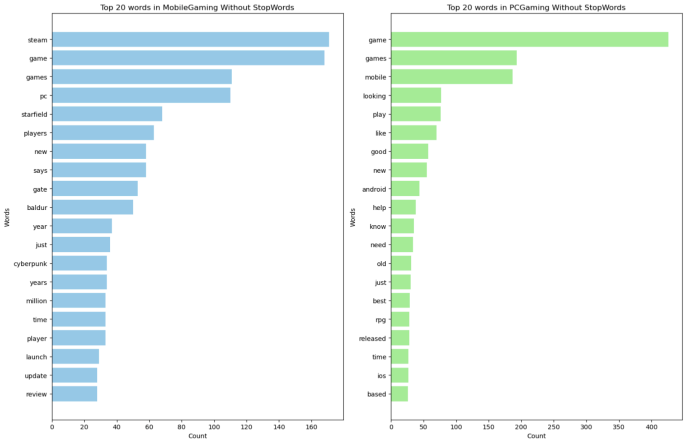


Figure 4

We standardized the text data using Python's re library to convert text to lowercase, remove special characters, emojis, and hyperlinks, and eliminate common stopwords. To understand the impact of stopwords, we created two dataframes for the word frequency analysis: one containing stopwords and one without stopwords. We applied Exploratory Data Analysis (EDA) and CountVectorizer to identify the top 20 words with the highest frequency in each subreddit. Figure 3 shows the word frequency analysis with stopwords, while Figure 4 shows the analysis without stopwords, revealing more context-specific terms. For sentiment analysis, we used the VADER sentiment analysis tool to provide sentiment scores for each title directly. Readability analysis was conducted using the `textstat` library to evaluate the complexity of the text, calculating metrics such as the Flesch-Kincaid readability score. The sentiment and readability scores were then used to compare the two subreddits. Figure 5 illustrates the distribution of sentiment and readability scores for the PCGaming and MobileGaming subreddits.

Data Analysis

Word Frequency Analysis

From Figures 3 and 4, it is evident that filtering out stopwords is crucial since the top words with the highest frequency were mostly stopwords when included. Using these words for analysis would not provide meaningful insights. Therefore, we redefined our analysis to focus on the most relevant words by filtering out common stopwords.

For statistical analysis, we combined the dataframes from both subreddits and used

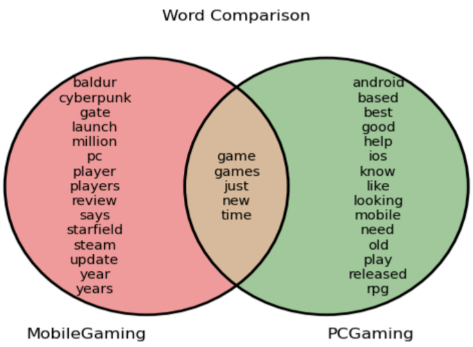


figure 5

CountVectorizer to convert the text into numerical data. The Chi-Square tests showed significant differences in word frequencies between the subreddits, with a p-value of 9.603111287063747e-300.

We also used various machine learning models to classify which subreddit a post belongs to based on word frequency. Surprisingly, models including stopwords performed slightly better than those excluding stopwords. This suggests that stopwords can still contribute to classification performance, highlighting the complexity of text data. However, the higher performance of models with stopwords might indicate overfitting, where the model learns the noise rather than the informative features. The results from the Logistic Regression, Random Forest, and Multinomial Naive Bayes models, as shown in the table, illustrate this finding. The Venn diagram in Figure 5 visualizes the overlap and distinctions in word usage between the two subreddits, emphasizing the importance of selecting meaningful words for analysis.

	Logistic Regression	Random Forest	MultinomialNB
Words counts	0.91111111111111	0.879629629629	0.9333333333333
Word counts without Stop Words	0.9240740740740	0.853703703703	0.9259259259259

figure 6

Sentiment Analysis

For the sentiment analysis, we utilized the VADER sentiment analysis tool to evaluate the sentiment scores for each title directly. VADER provides a sentiment score that ranges from -1 to 1, where negative values indicate negative sentiment, positive values indicate positive sentiment, and scores around 0 indicate neutral sentiment.

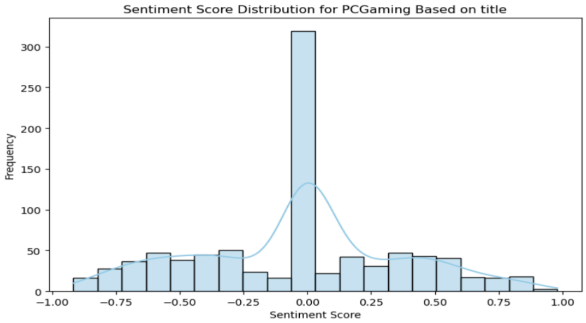


figure 7

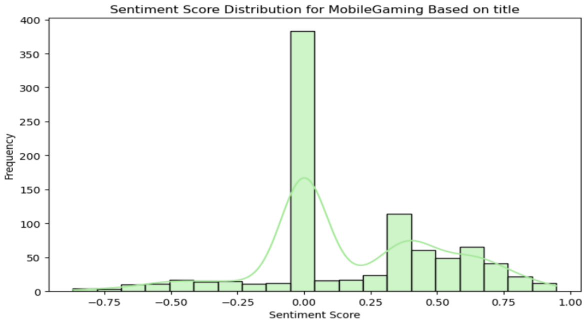


figure 8

To statistically compare the sentiment scores between the two subreddits, we applied the Mann-Whitney U test, which is suitable for comparing two independent samples, especially when the data does not follow a normal distribution. The results of the Mann-Whitney U test indicated a significant difference in sentiment scores between the PCGaming and MobileGaming subreddits, with a p-value of 2.9937340274256154e-29. The sentiment score distributions for both subreddits are shown in Figure 7 and 8, which illustrate the differences in sentiment expressed in the post titles. Although we attempted to use sentiment scores to classify the subreddit of the posts, we encountered an error when using the Multinomial Naive Bayes classifier. The sentiment_title data might not provide sufficient discriminatory power for the classifier because it could be too simplistic to capture the complexities of the subreddit content.

	Logistic Regression	Random Forest	MultinomialNB
Sentiment Score	0.5944444444444	0.635185185185	/

figure 9

These findings suggest that sentiment analysis helps understand the emotional tone in different gaming communities on Reddit and can distinguish between subreddits based on post titles. However, sentiment scores alone might be too simplistic for effective classification using machine learning models.

Readability Analysis

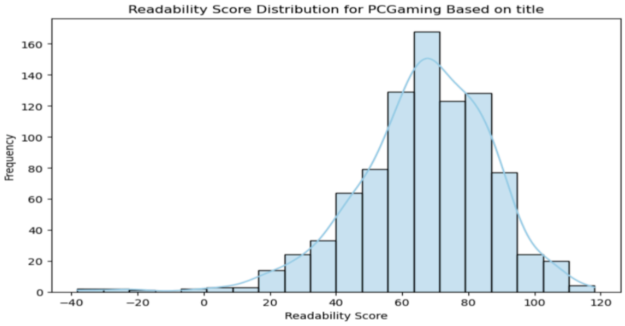


figure 10

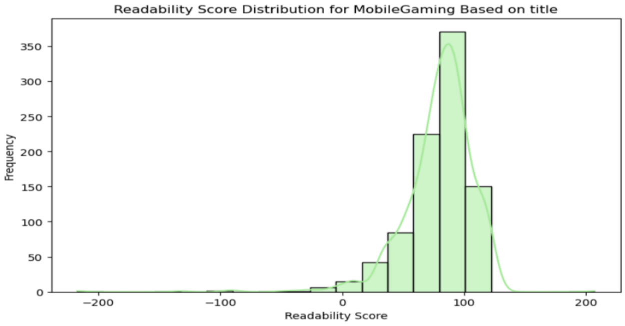


figure 11

For the readability analysis, we applied the same methodology as the sentiment analysis. We utilized the “textstat” library to calculate the readability scores of the titles, specifically using the Flesch-Kincaid readability score, which typically ranges from 0 to 100, with higher scores indicating easier readability. The readability score distributions for both subreddits are shown in Figure 10 ans 11. To statistically compare the readability scores between the two subreddits, we applied the Mann-Whitney U test, which is suitable for comparing two independent samples. The results of the Mann-Whitney U test indicated no significant difference in readability scores between the PCGaming and MobileGaming subreddits, with a p-value of 1.0. This suggests that the readability levels of the titles in both subreddits are quite similar. The machine learning models, Logistic Regression and Random Forest, were used to classify the data based on readability scores, as shown in the table below. However, the performance of these models indicates that readability score alone might not be a strong discriminator for classifying subreddit content.

	Logistic Regression	Random Forest	MultinomialNB
Readability Score	0.6537037037037	0.6796296296	/

figure 12

Limitation

The project faced several limitations that affected the scope and accuracy of the findings. Firstly, the dataset was limited to the top 1000 posts from each subreddit ("PCGaming" and

"MobileGaming"), which may not fully represent the entire spectrum of discussions and sentiments within these communities. Additionally, the VADER Sentiment Analyzer, while effective for general sentiment analysis, struggled with detecting nuanced expressions such as sarcasm and irony, potentially leading to inaccurate results. The readability analysis was also limited, focusing on basic metrics and ignoring multimedia content that could influence user engagement. Moreover, the analysis only used the data from the titles of posts, as the self-text and submission text were often missing or less unique, which might have provided a more comprehensive view.

Conclusion

This project analyzed linguistic characteristics to distinguish posts in the PCGaming subreddit from those in the MobileGaming subreddit. Filtering out stopwords was crucial for meaningful word frequency analysis, as including them did not provide insightful results. Chi-Square tests showed significant differences in word frequencies, with word counts without stopwords yielding the most relevant insights. Using these refined word counts for machine learning models, Logistic Regression achieved the highest accuracy at 0.924. Sentiment analysis revealed distinct emotional tones, with PCGaming posts tending towards neutral sentiment and MobileGaming posts showing a broader range of sentiments. However, sentiment scores alone were not highly effective for classification. Readability scores were similar across both subreddits and were not strong classifiers. Overall, word frequency analysis was the strongest indicator for distinguishing subreddit content, and filtering out stopwords was essential for improved performance and relevance.

Accomplishment Statement:

Yinxuan Zhang:

- Successfully led the data collection and cleaning phases of a data analysis project, distinguishing linguistic characteristics in the PCGaming and MobileGaming subreddits.
- Extracted a robust dataset of 900 posts per subreddit using PRAW and Python, ensuring high data quality by removing duplicates and irrelevant content.
- Conducted detailed data processing and standardization, converting text to lowercase, and removing special characters and common stopwords, facilitating accurate analysis.
- Collaborated on the application of Exploratory Data Analysis (EDA) and CountVectorizer to identify top word frequencies, revealing distinct linguistic patterns between subreddits.
- Presented initial findings and cleaned data for further analysis, contributing significantly to the project's success.

Yanjun Qian:

- Effectively led the data analysis and reporting phases of a data analysis project, distinguishing linguistic characteristics in the PCGaming and MobileGaming subreddits.
- Applied Exploratory Data Analysis (EDA) and CountVectorizer to identify top word frequencies, and utilized VADER sentiment analysis to evaluate emotional tones in posts.
- Conducted statistical analysis using the Mann-Whitney U test to compare sentiment scores between subreddits, revealing significant differences.
- Applied the textstat library to calculate readability scores of post titles, comparing readability levels between the two subreddits.
- Achieved a 92.4% classification accuracy with Logistic Regression by refining word counts, demonstrating significant differences between subreddits with a Chi-Square test p-value of 9.60e-300.
- Presented comprehensive reports and visualizations, highlighting key insights and contributing to the understanding of how platform-specific factors influence online discussions.