From Streams to Review:

Analyzing Consumer Feedback on Spotify

Armor Cao, Becky Song, Liujun Chen

Table of Contents

01. Research Question

02. Comparative Analysis

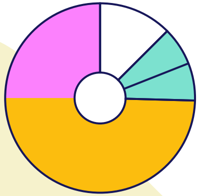
03. Sentiment Analysis

04. Classification

Research Question

- **We developed our research questions by creating three sub-questions to zoom in and conduct a deep investigation.**

- **Research Question:** How can Spotify Play Store reviews be analyzed to understand evolving user satisfaction, brand perception, and review patterns?
  - **Sub-1:** What are the most common words discussed in user reviews?
  - **Sub-2:** What is the overall sentiment trend of Spotify reviews?
  - **Sub-2:** Can review ratings be predicted based on review text features?



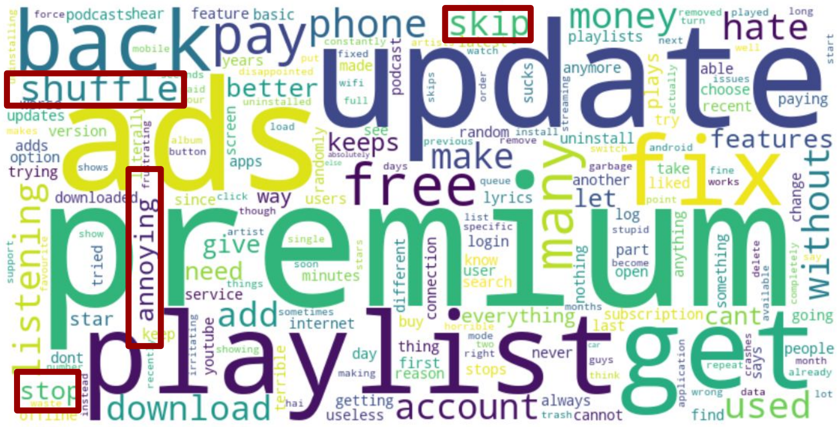
01.

Comparative Analysis

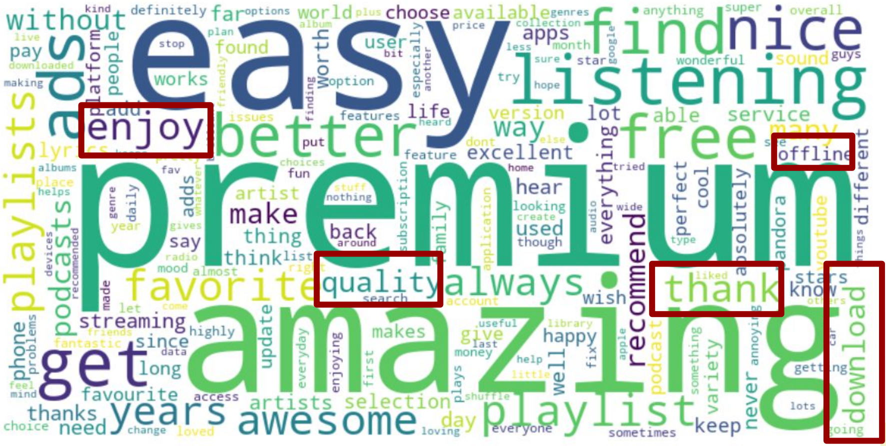
Word Cloud by Review Rating

reviews with > = 7 words, removed stop words

Review = 1



Review = 5

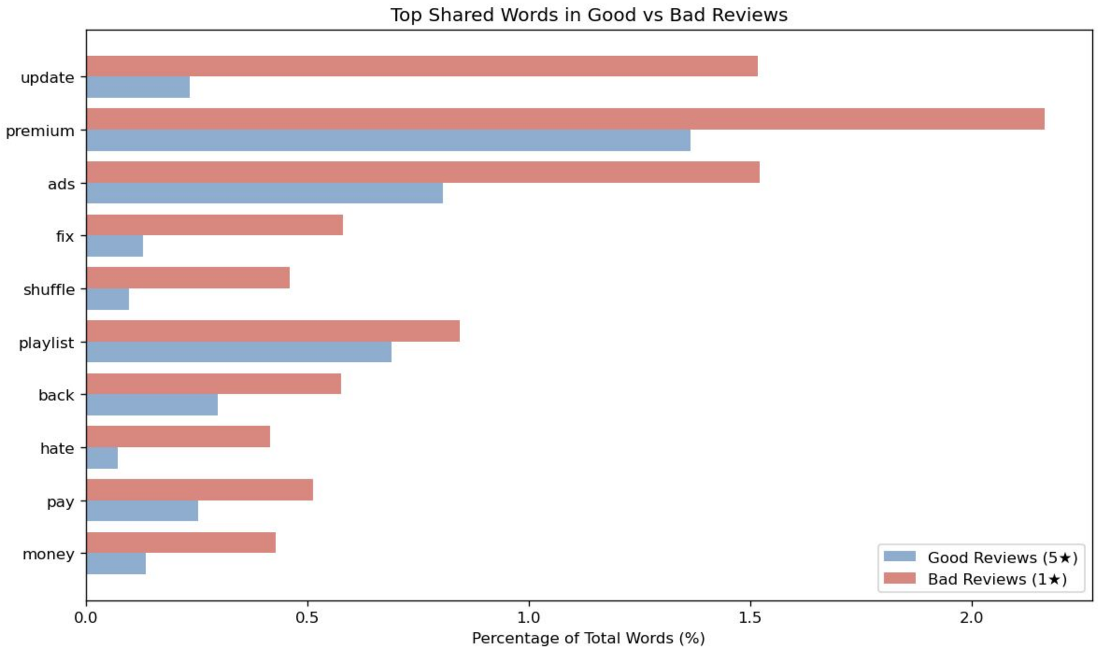


TOP 5:Premium, ads,update,playlist,fix

TOP 5:Premium, Easy, amazing, listening,ads

TOP 10 Common Words in 1 and 5 Reviews

- Users care about **updates, premium pricing, ads, and app performance**.
- These terms appear more frequently in 1-star reviews, suggesting that negative experiences cluster around these shared concerns.
- **Positive example**: ad-free Premium, good playlist variety, generally smooth usage
- **Negative example**: crashes, playback interruptions, loading delays, and shuffle unpredictability



02.

Sentiment Analysis

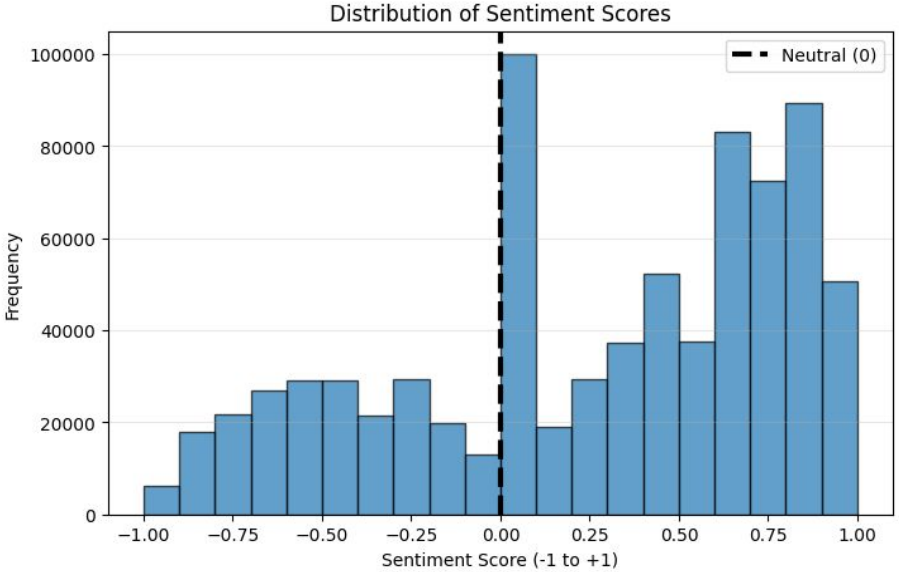
Sentiment Score Distribution

Positively Skewed

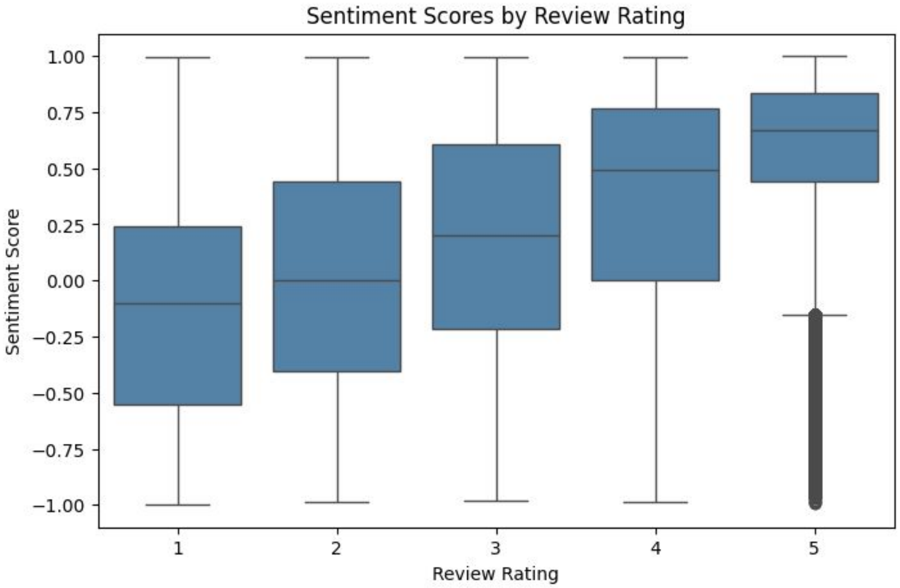
- The largest cluster is near-neutral (0).
- The second-largest group is strongly positive (0.6-0.9).

Negatives are Scarce

- The entire negative range is significantly smaller than the positive range.
- 75% of all scores are -0.12 or higher.



Sentiment Score vs. Review Rating



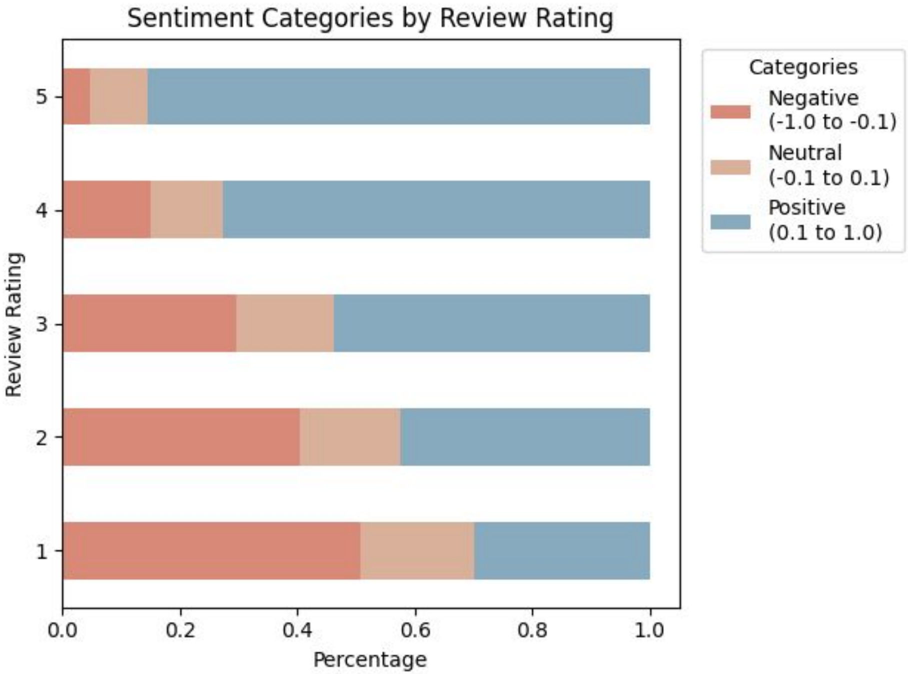
Strong Correlation

- As the user's Review Rating increases, the Sentiment Score also clearly increases.

The 5-Star Anomaly

- A large group of 5-star reviews contains text with highly negative sentiment scores.
- Explanation: people notes their frustrations but still give 5 points-”Its really good tbh, the ads get a little annoying……”

Sentiment Category vs. Review Rating



Progressive Trend

- The share of negative text reviews consistently shrinks as the rating increases from 1 to 5.

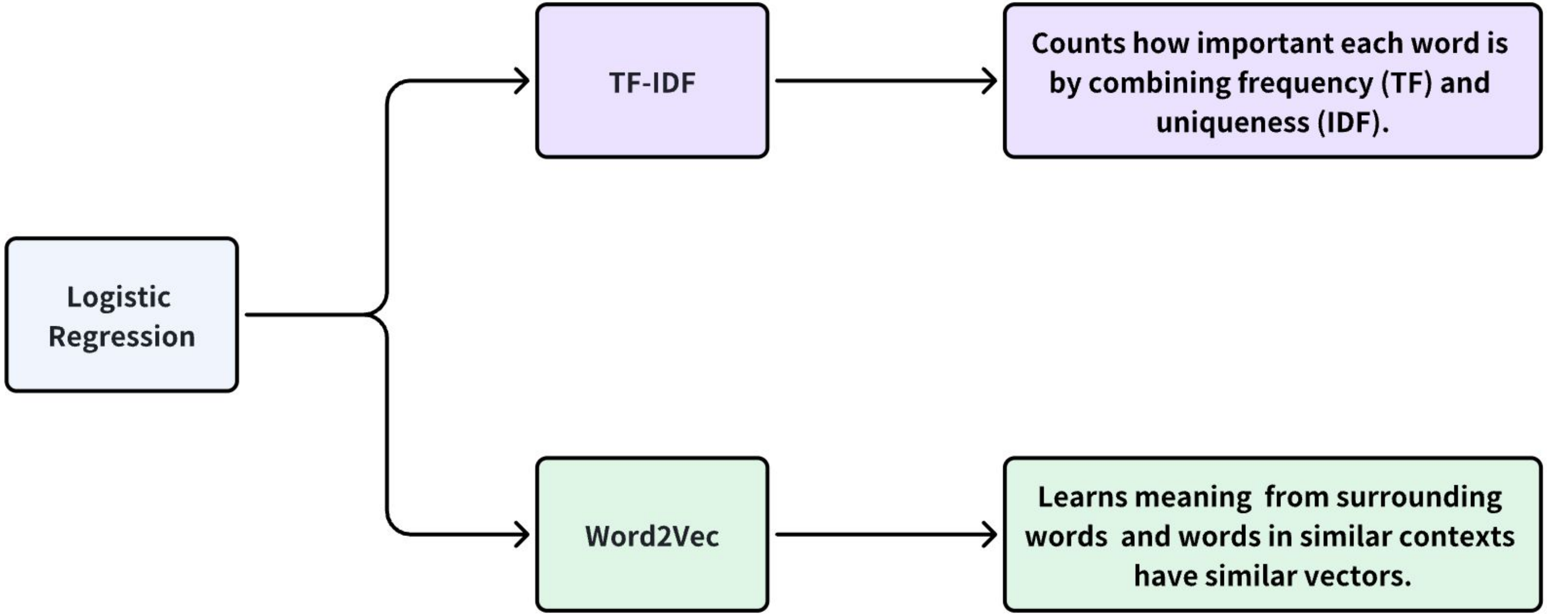
Mixed Negative Signals

- 1-star reviews is highly conflicted, containing only ~50% negative text reviews, but 30% of reviews with positive sentiment score.

03.

Classification

Classification Pipeline



Logistic Regression

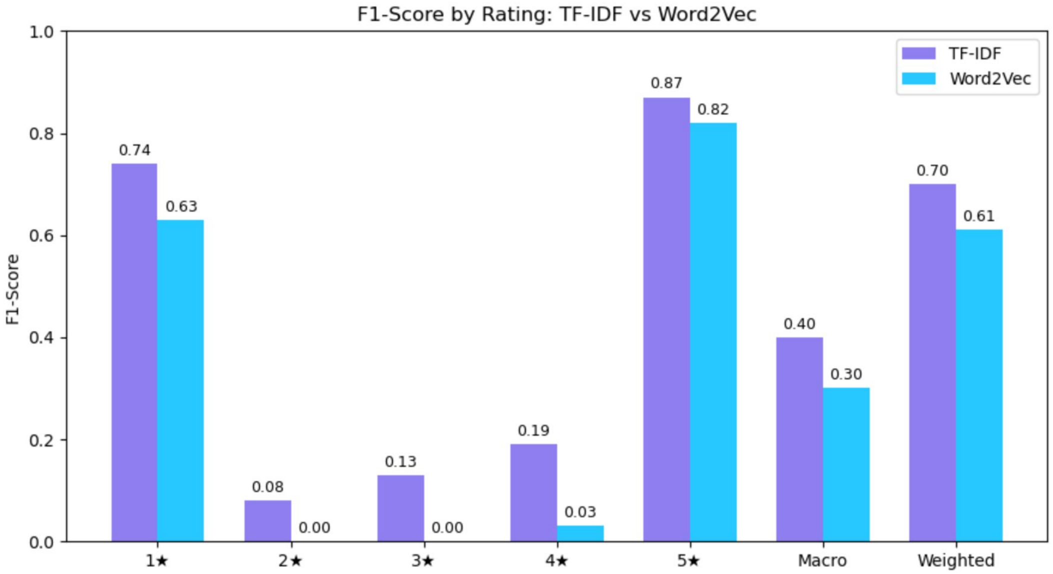
TF-IDF

Counts how important each word is by combining frequency (TF) and uniqueness (IDF).

Word2Vec

Learns meaning from surrounding words and words in similar contexts have similar vectors.

F-1 Score Comparison



TF-IDF

- Performs strongly on **extreme ratings (1 star & 5 star)**, so it has good ability to classify clear positive or negative sentiments.
- **Mid-range ratings** show **low F-1 scores**, meaning it has difficulty predicting neural or mixed tones.
- Macro F1 = 0.40, meaning the performance is **uneven** across classes, strong only for extremes.
- Weighted Avg = 0.70, meaning the overall performance driven by **majority classes.**

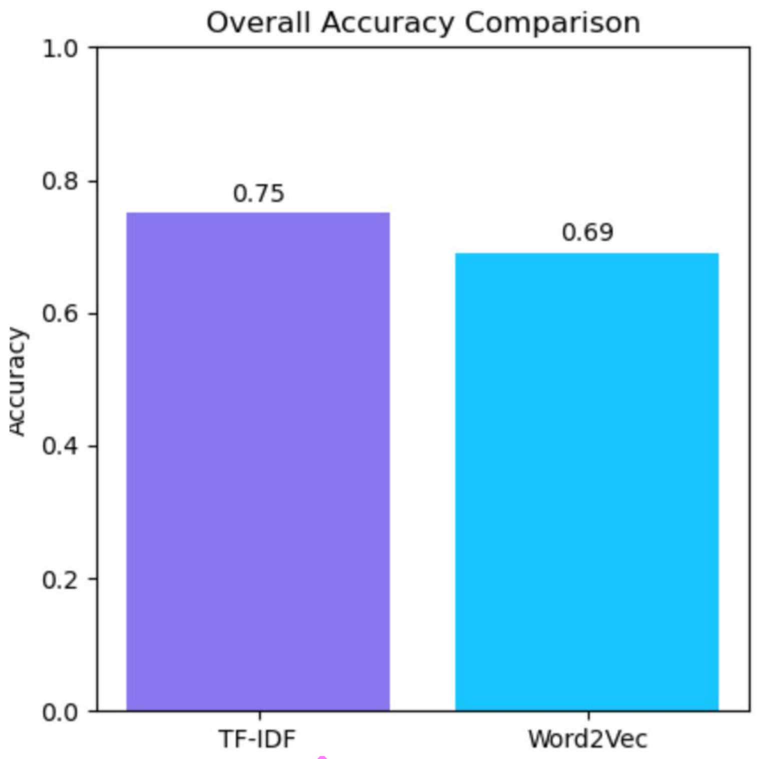
Word2Vec

- Also performs well on 5 star reviews, but weaker on 1 star reviews compared to TF-IDF.
- F-1 scores are near **0 for 2 start - 4 star** reviews, meaning that it **fails** to make accurate predictions for nuanced or balanced language.
- Macro F1 = 0.30 and Weighted Avg = 0.61, so overall it has **weaker and less balanced classification**.

Overall

- TF-IDF **outperforms** Word2Vec across all F-1 metrics.
- Both models perform best for extreme sentiments but fail to capture middle range reviews.

Accuracy Comparison



TF-IDF

- Accuracy = 0.75, meaning the model predicts about 75% of the reviews correctly, so it has overall strong classification ability.
- So even though mid-range classes have lower F-1, overall accuracy is still high due to the correct predictions in extremes.

Word2Vec

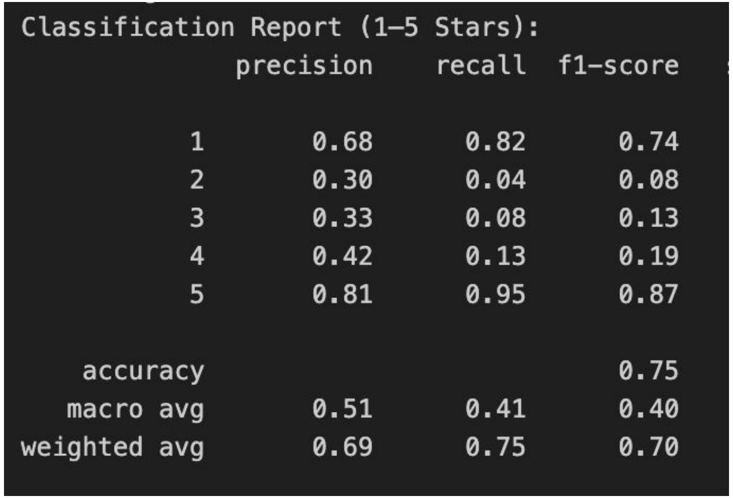
- Accuracy = 0.69, so it is slightly weaker at predicting correct ratings compared to TF-IDF.
- It has the same issue as TF-IDF that the overall accuracy was not low, but mainly drive up by the extremes. There are still many misclassifications for neutral or mixed reviews from 2 star to 4 star.

Overall

- TF-IDF achieves higher overall accuracy, confirming its superior predictive power in this task.
- Both models can identify clear sentiment polarity (positive/negative), but TF-IDF does it more accurately and consistently.

Original Results

TF-IDF

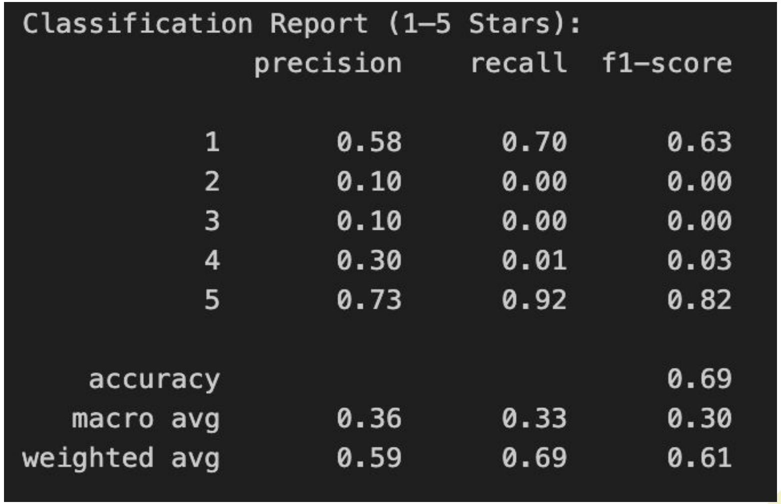


```
Classification Report (1–5 Stars):
              precision    recall  f1-score

           1       0.68      0.82      0.74
           2       0.30      0.04      0.08
           3       0.33      0.08      0.13
           4       0.42      0.13      0.19
           5       0.81      0.95      0.87

    accuracy                           0.75
   macro avg       0.51      0.41      0.40
weighted avg       0.69      0.75      0.70
```

Word2Vec



```
Classification Report (1–5 Stars):
              precision    recall  f1-score

           1       0.58      0.70      0.63
           2       0.10      0.00      0.00
           3       0.10      0.00      0.00
           4       0.30      0.01      0.03
           5       0.73      0.92      0.82

    accuracy                           0.69
   macro avg       0.36      0.33      0.30
weighted avg       0.59      0.69      0.61
```

Thanks!