What Game Are You Playing?

Project 3 - Karen Chien

Project: Classifying different subreddits

6000 total subreddit posts and comments collected from two game subreddits

Data cleaned and processed to create multiple classification models

Which model performed best based on accuracy?

Subreddit Data

Stardew Valley

- Dating and farm simulator game
- Released in 2016
- Sold over 15 million copies by 2021

vs.

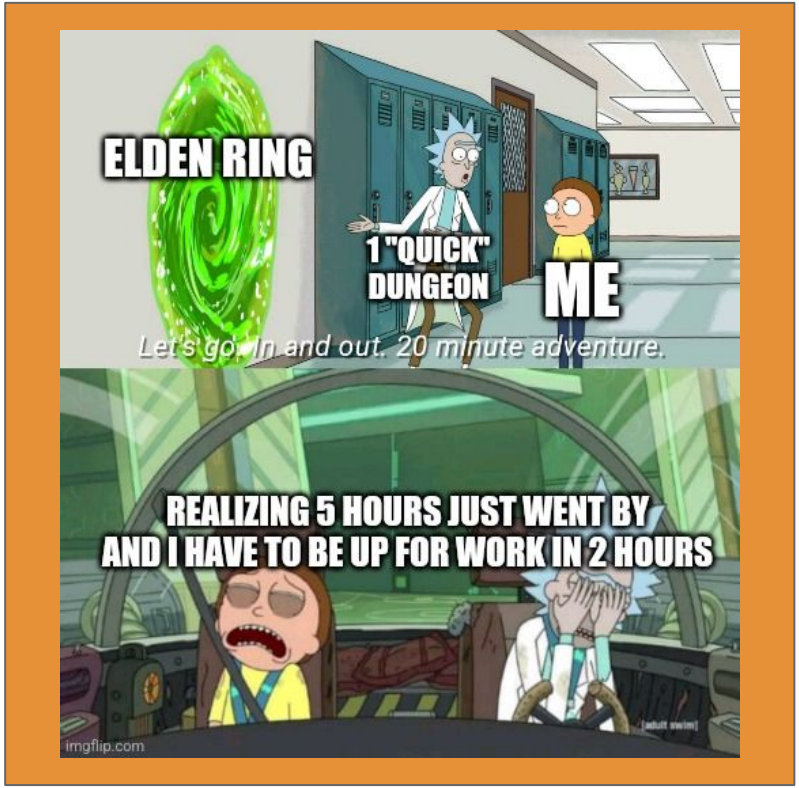
Elden Ring

- Action RPG game
- Released in 2022
- 12 million copies sold three weeks after release date

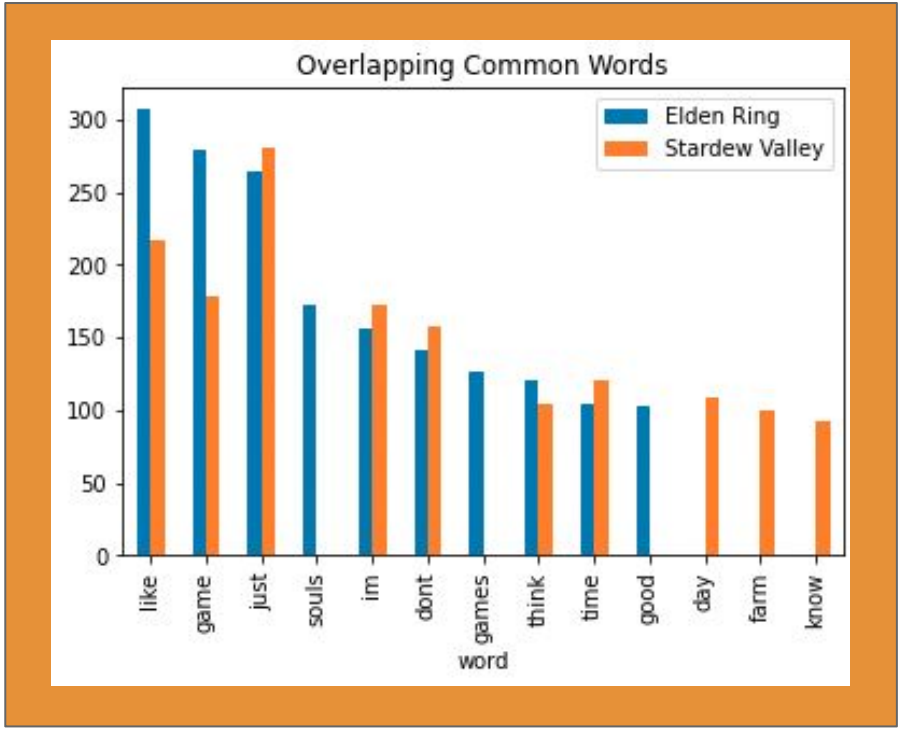
Subreddit Data Challenges

- **Posts contained memes, screenshots, or photos** – Post titles were combined with the post body to get more textual information.

- **Additional comment data was extracted –** 66% of dataset taken from comments.



Subreddits' EDA Observations



- **Overlapping words** – Top 10 common words had overlap that were not caught by stopword.
- **Distinct words related to gameplay –** No overlap and could be helpful classifiers

Optimizing During Pre-Processing

Strategy 1: Created feature if post contained words related to distinct gameplay

RESULT: Too many features in the model, minimal impact

Strategy 2: Created list of overlapping words that occurred above the median in each subreddit and added to stopwords.

RESULT: Progress in improving accuracy for some models and reduced variance overall

Optimizing for Accuracy - Results

Model	Best Train Score	Best Test Score
Baseline	51.0	51.0
Logistic Regression	92.0	82.0
KNN Classifier	95.0	64.0
Multinomial Naive Bayes	91.0	82.0
Random Forest Classifier	95.0	80.0

Best Model - Multinomial Naive Bayes

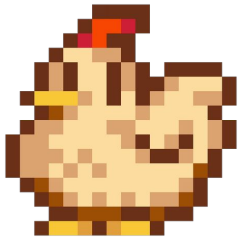
Model	Best Train Score	Best Test Score
Baseline	51.0	51.0
Logistic Regression	92.0	82.0
KNN Classifier	95.0	64.0
Multinomial Naive Bayes	91.0	82.0
Random Forest Classifier	95.0	80.0

Follow-Ups

- **Deeper dive into overlapping words** – What can be removed from custom stopword list? What can be added?

- **Bring in more comment data, or use strictly comments –** Since posts produced minimal text analysis, can subreddit comments provide additional boost?

- **Highlight gameplay-specific words –** TFIDFVectorizer did not produce a higher-performing model. Are there other ways to weight words related to unique gameplay?



Questions? Feedback?