Project 3 - Web APIs & NLP
Console Game Subreddit

Seth, Jervin, Lucas

Problem Statement

A gaming store has newly set up around the corner. Its owners wish to build its business activities, and increase physical and online presences.

To start off, they would like to find out which major consoles (PS5 and Xbox) are trending. At the same time, they are exploring the idea of having a forum on their website that allows for discussion among gaming community, and e-commerce section to include product reviews.

We were engaged to develop a classification model that predicts which category a post belongs to. This will be helpful for their forum moderation/upkeep, and users will also experience better navigation of the store's online space with the help of accurate tagging/sections allocation.

We are also tasked to identify the recent topics of interest and the community's sentiments towards them.

Key Questions

- Which community is more active?
- What are the trending topics for each community?
- Which products should we focus our marketing on?
- Regarding top topics, what are the community's sentiments and emotions towards them?
- What is the best model to classify post?

Process

- Data Collection
- Data Cleaning and Exploration
- Preprocessing
- Modelling
- Model Evaluation
- Sentiments and Emotions Analysis

Data Collection and Cleaning

Data Collection

Web scraped using Pushshift Reddit API:

- Subreddits: PS5 (*PlayStation 5*) and XBox Series X
- Posts extracted: Submissions, excluding comments
- Number of posts: 15,000 each
- Time frame: Before 24 Jun 2022 0000hrs
- Data used:
  - Title
  - Selftext (posts' text content)
  - Created_UTC (date/time of post creation)

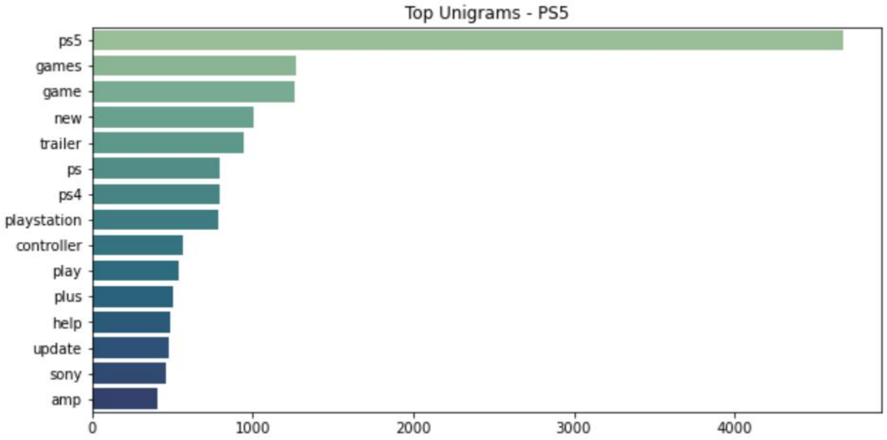
Data Cleaning

- Null values
- Selftext column
    - Significant amount with non-constructive text; e.g. [removed] and [deleted]
- Remove non-English characters, symbols, HTML links, stopwords
- Remove rows with duplicate titles
- Convert emojis to words using demojize
- Remove documents with less than 2 words

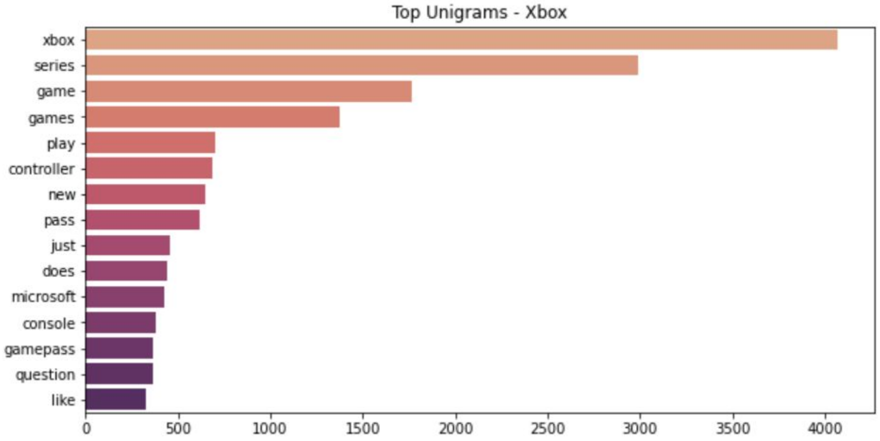
Exploratory Data Analysis

Unigrams

- Top occurring unigrams for both datasets contain some common words like ‘game’, ‘games’, ‘controller’ and ‘play’.
- Includes company name, console name and subscription service.



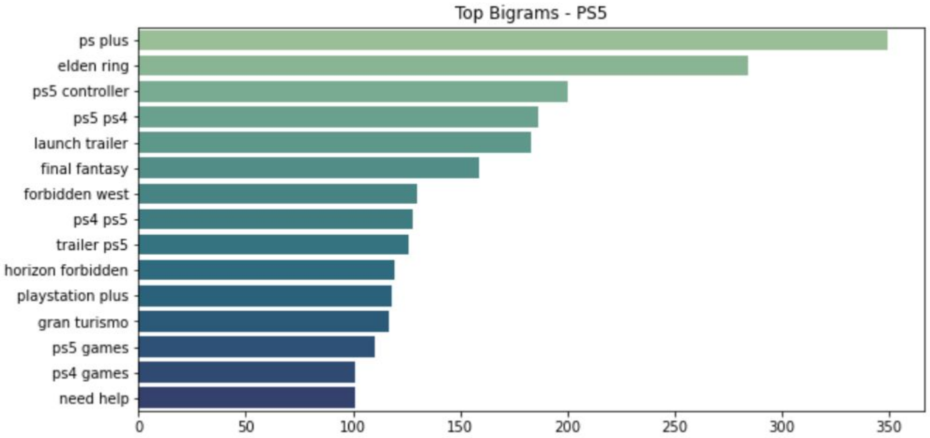
Top Unigrams - PS5

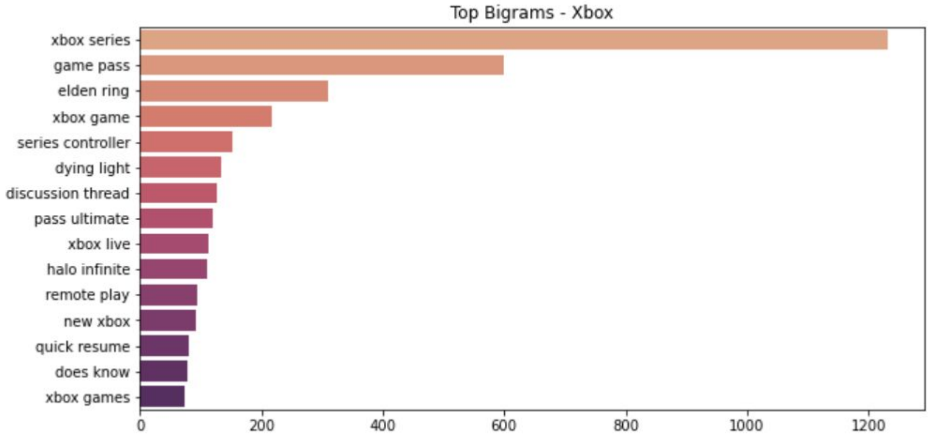


Top Unigrams - Xbox

Bigrams

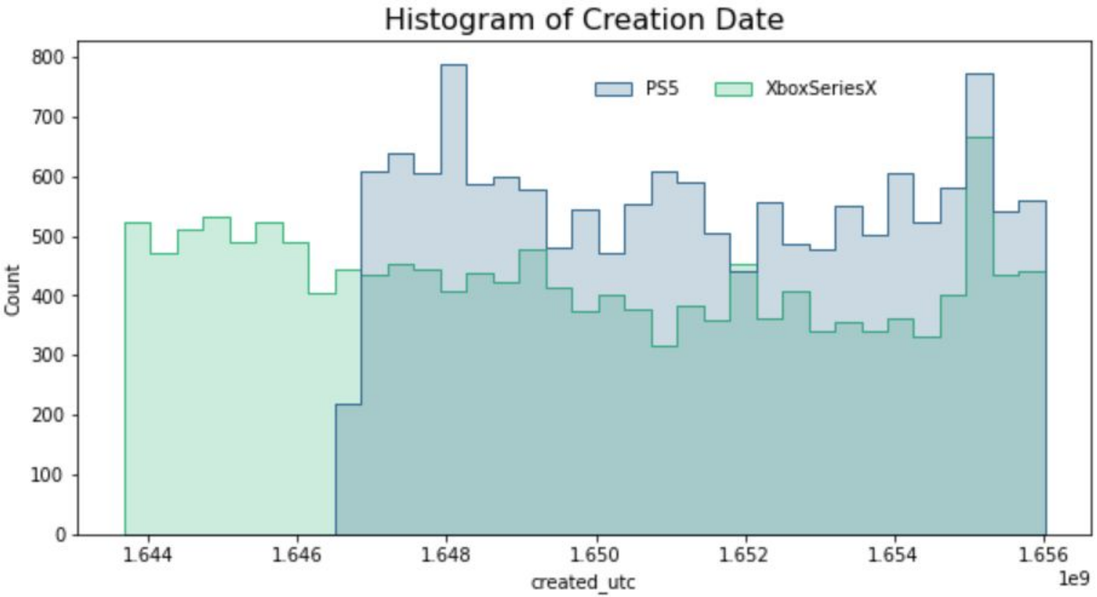
- Top bigrams charts shown the most mentioned games of each console.
- ‘Elden Ring’ is the only game that appears in both bar plots, and is also the most mentioned game in both subreddits.





Post Creation Date

- Earliest post for:
  - PS5: 8 Mar 2022
  - Xbox Series X: 31 Jan 2022
- No. of posts created per day on the PS5 subreddit is roughly 20-40% on average higher than Xbox.



Preprocessing and Modelling

Preprocessing flow

- Remove URL, non-english characters
- Tokenization - Separate sentences into individual word
- Remove stopwords - Customized list
    - Custom stopwords - PS5, Xbox Series X
- Stemming/Lemmatization
    - Number of columns after Stemming: 11354
    - Number of columns after Lemmatizing: 13581
- CountVectorizer/TF-IDF
    - Chosen TF-IDF, it factors in the relevance of the word relative to the corpus

Classification model metrics

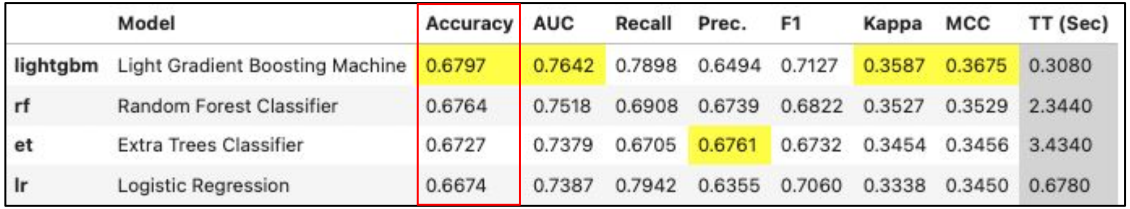
- Consideration: Accuracy, Recall, Precision, F1

- Chosen metric: Accuracy
    - Target variable is balanced (50% each)
    - Consequences of False Negative and False Positive is more or less the same
    - No prioritizing of True Positive/Negative or False Positive/Negative required in this business case

Modeling

- Baseline: Naive bayes

- Top 3 models
  - Light GBM
  - Random Forest
  - Logistic Regression



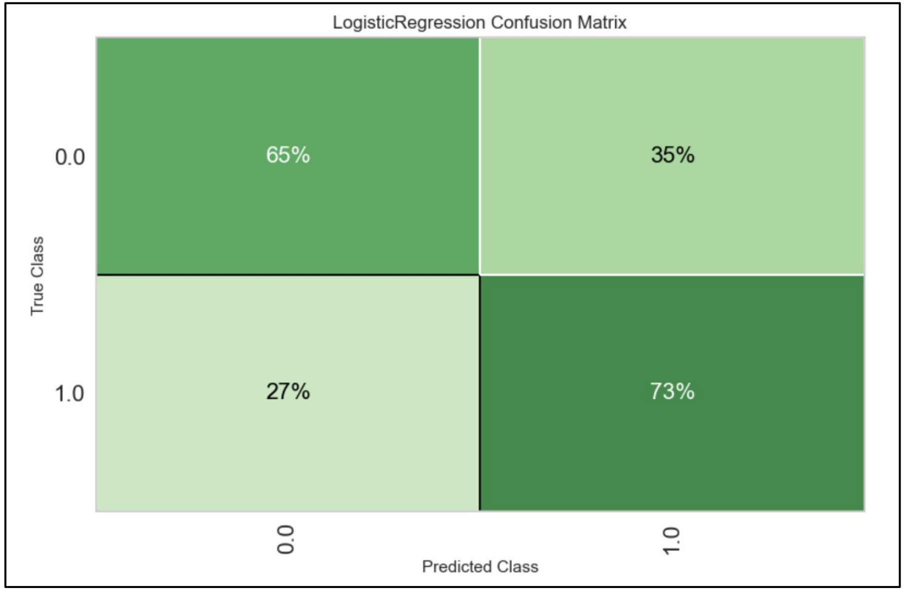
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.6797	0.7642	0.7898	0.6494	0.7127	0.3587	0.3675	0.3080
rf	Random Forest Classifier	0.6764	0.7518	0.6908	0.6739	0.6822	0.3527	0.3529	2.3440
et	Extra Trees Classifier	0.6727	0.7379	0.6705	0.6761	0.6732	0.3454	0.3456	3.4340
lr	Logistic Regression	0.6674	0.7387	0.7942	0.6355	0.7060	0.3338	0.3450	0.6780

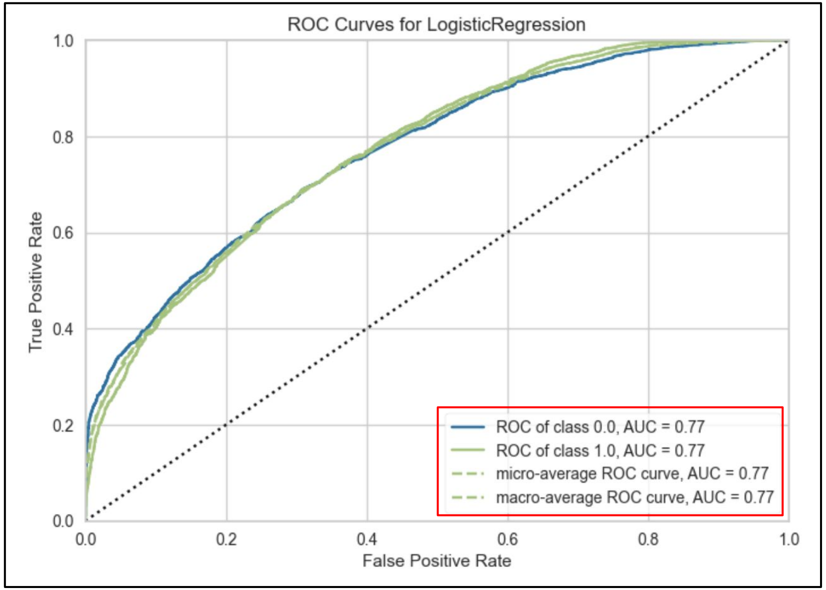
Chosen model and evaluation

- Logistic Regression to classify whether a post is relating to PS5 or Xbox
    - Highest Accuracy score and well-fitted
    - Interpretable
    - Computationally cheaper / faster
    - Appropriate for a binary classification case

Finalised model (hyper-parameter tuned)				
Model	Logistic Regression	Random Forest	LightGBM	Naive Bayes
Train Score	0.721	0.949	0.747	0.714
Test Score (Holdout)	0.703	0.689	0.697	0.700

Model Scores (EDA)

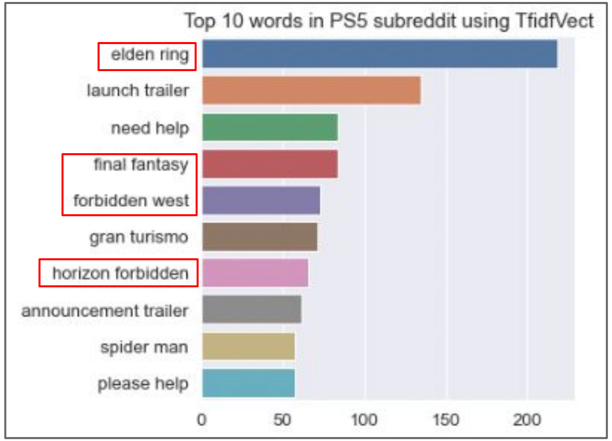


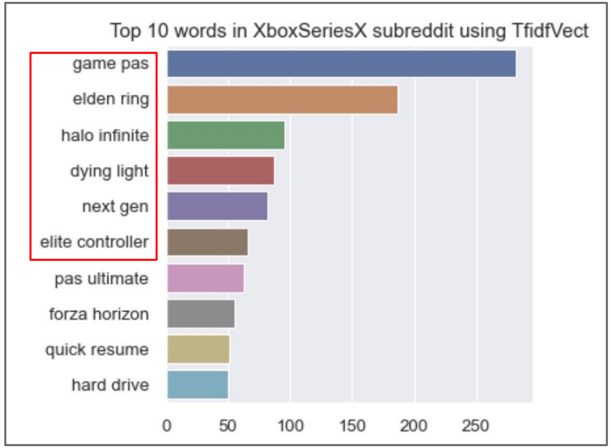


Sentiment and Emotion Analysis

Analysis on Hot Topics:

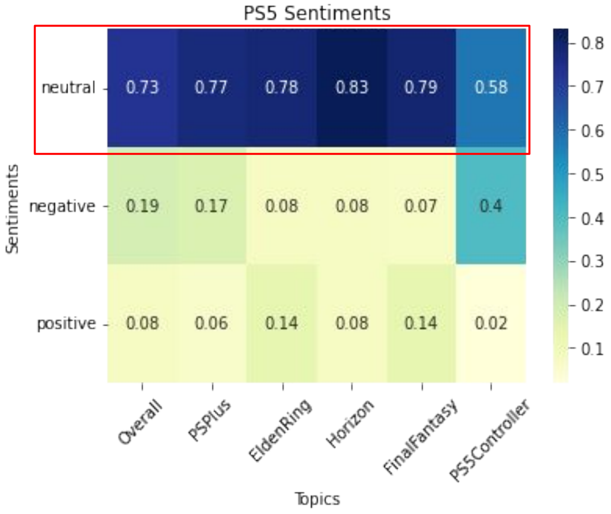
- PS5
  - PS5 Plus
  - PS5 Controller
  - Elden Ring
  - Horizon Forbidden West
  - Final Fantasy
- Xbox
  - Game Pass
  - Elden Ring
  - Dying Light
  - Halo Infinite
  - Xbox Series Controller

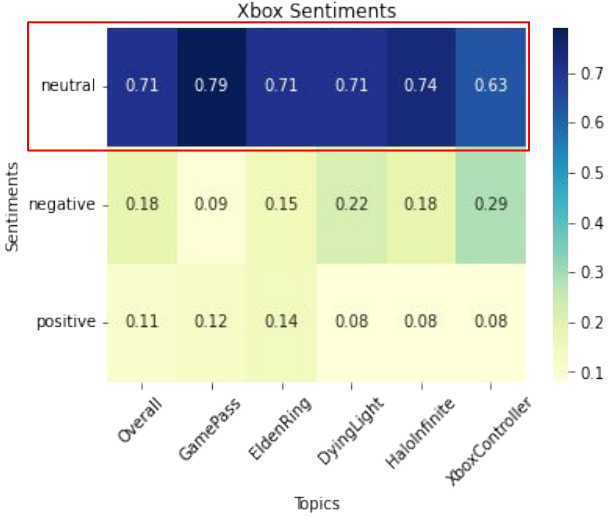




Sentiments

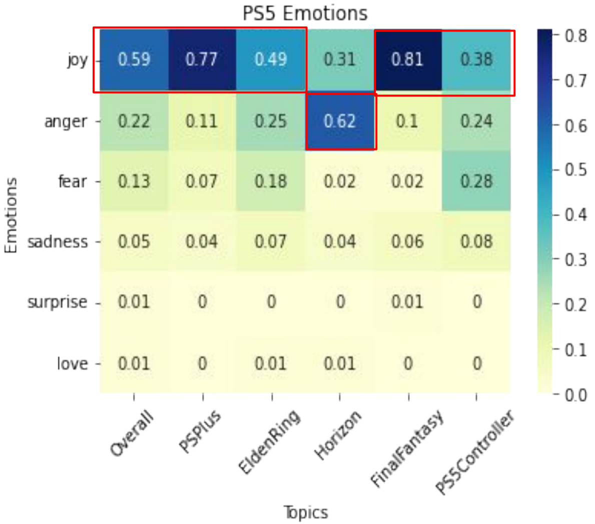
- Majority of Post seem to be neutral
  - Questions, news, trailers
- Post relating to the top games are more positive for both console
  - Dying Light is the exception - possibly due to its name
- Topics relating to controllers are more negative
  - Complaint
  - Issues
- Subscription Service
  - Game Pass is more well-received than PS Plus

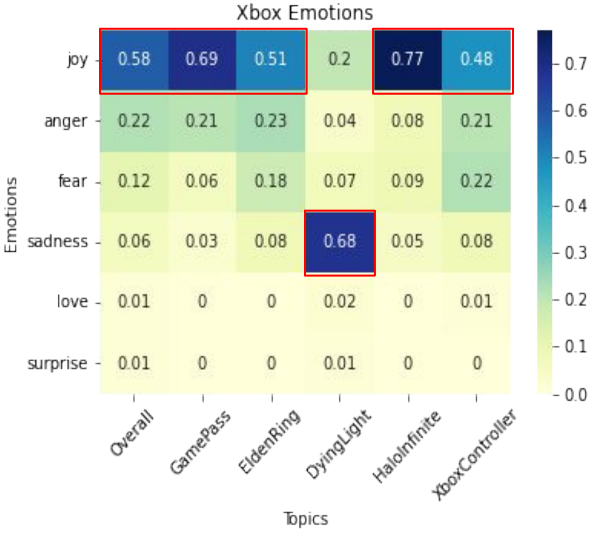




Emotions

- Joy is the dominant emotion
- Some odd results
  - Anger for Horizon Forbidden West
  - Sadness for Dying Light
- Contradicts sentiments analysis
  - PS Plus and controllers - ‘Joy’
- Naming of the game could be very influential
  - ‘Dying’, ‘Fantasy’
- Results not accurate
  - Neutral nature of most post (news/questions)





Insights and Recommendations

Insights

- PS5 Community is the more active of the two
- Hot topics for each subreddit are their Top games, controllers and Subscription Service
- Of the trending topics, post relating to the top games are the most well-received
- Both Xbox and PS5 topics have many similar words

Recommendations

To improve accuracy and modelling:

- Collect more data - Twitter, Facebook
  - More information that could distinguish the two topics
- Further customize stopwords list
  - Further analysis on the subject matter to remove words
- Utilising GPU/cloud service speed up processing of bigger data volume

Business

- Game store to focus more on products identified:
  - PS5
  - Game titles: Elden Ring, Final Fantasy, Halo Infinite, Dying Light
  - Avoid Controllers of each console
- Logistic Regression classification model may possibility developed into automated category tagging system to automate business processes.

Limitations

- Model may not be able to classify well when both subreddits contain similar top few words (i.e. game)
- Time and computing constraints
  - Time needed to complete processing limits hyper-parameters tuning/unsupervised learning
- Although Logistic regression is easily interpretable, it has some assumptions
  - Linearity between the features and the logit of the probability when Y=1
  - Features are independent of each other
- Emotion/Sentiments analysis may not be well-interpreted when texts are dominated by game names and interpreted in word meaning context (*e.g. "Dying Light"* = sad/negative)