Classifying Reddit Posts

An analysis of PS5 and XboxSeriesX sub-reddits

By Chee Howe, Hong Yee, Benjamin, Daiyu

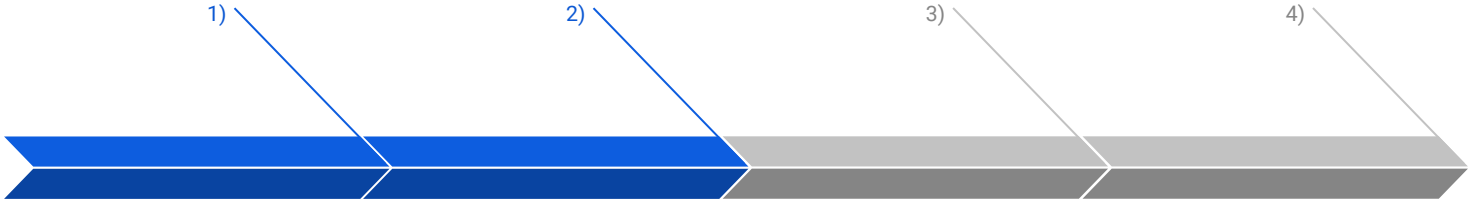
Problem Statement

Seeking a contingency plan in the event of a server glitch, Reddit is looking to develop a model to sort posts into the respective subreddits when subreddit posts get mixed up. As a proof of concept, Reddit has requested for us to;

1. Create a model that best classify which subreddit a post came from between 2 subreddits.

2. Identify words that will most differentiate our two chosen subreddits.

Methodology



Data Collection

Scrap posts from PS5 and XboxseriesX subreddits through Reddit's API.

Data Cleaning and Preprocessing

Retained only the 'selftext', 'title' and 'subreddit' columns.

Cleaned and prepared the data for modelling.

Model Selection and Analysis

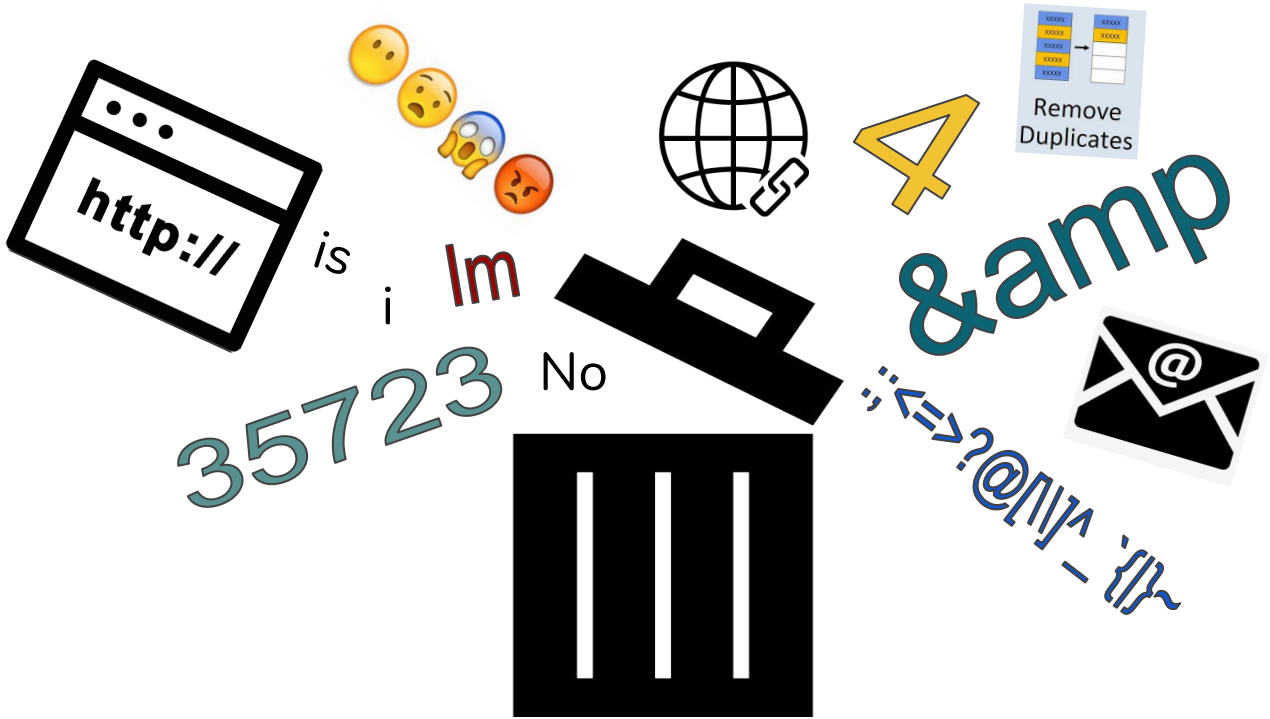
Selected best performing model based on highest **accuracy score**.

Identified the top 30 words that influence the classification model.

Conclusion

Recommendations to improve our model and suggestions for way ahead.

Data Cleaning using regu[ular] expr[essio]n



Impute missing values

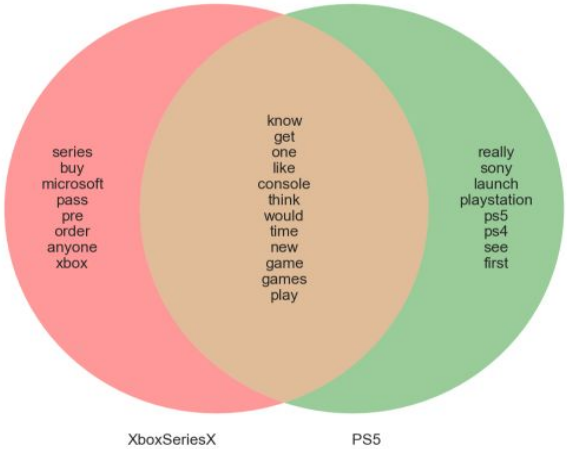
‘title’ + ‘selftext’ = ‘title selftext’

‘title’ + ‘ ’ = ‘title’

	title	selftext	combined_text
0	random fan decided make this awesome xgp trailer credit vincenzo fayez channel		random fan decided make this awesome xgp trailer credit vincenzo fayez channel
1	the taco bell deal available country	pretty sure there taco bell here saudi arabia but idk the prize thing available will very appreciated can help yes checked google didn say anything	the taco bell deal available country pretty sure there taco bell here saudi arabia but idk the prize thing available will very appreciated can help yes checked google didn say anything
2	state mouse and keyboard support xbox consoles	buying series regardless mouse and keyboard support but was wondering what the current state mainly play shooters and when playing console with controller feels like drunk its just incredibly slow and inaccurate from what understand some games support mouse and keyboard single player that true for most shooters don really care about multiplayer games anymore but would like play titles like doom with mouse and keyboard that possible	state mouse and keyboard support xbox consoles buying series regardless mouse and keyboard support but was wondering what the current state mainly play shooters and when playing console with controller feels like drunk its just incredibly slow and inaccurate from what understand some games support mouse and keyboard single player that true for most shooters don really care about multiplayer games anymore but would like play titles like doom with mouse and keyboard that possible

Remove stop and overlapped words

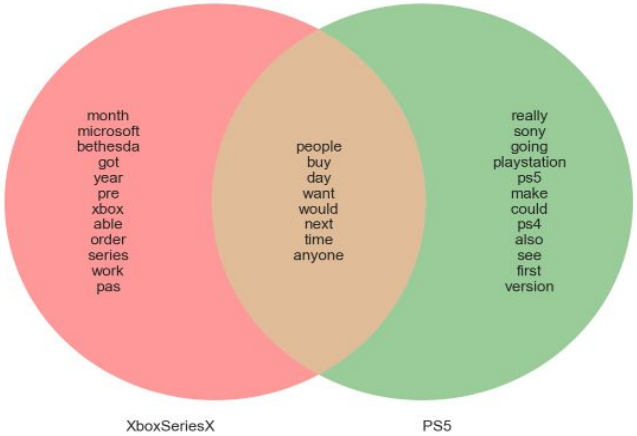
after ‘english’ stopwords removal



Top 40

high frequency overlapping words removal

after remove high freq overlapped words



Cleaned Top 40

Exploratory Data Analysis

Both subreddits are home video gaming consoles, going to be launched this November at the same price. They are similar in many ways and often being compared against.

‘xbox’ appeared 1023 times, ‘series’ appeared 742 times, ‘microsoft’ appeared 239 times.
‘ps5’ appeared 888 times, ‘ps4’ appeared 392 times, ‘sony’ appears 320 times.

- Top predictive unique words: product name and manufacturer’s name.
- Top overlapped words: ‘game’, ‘new’, ‘play’, ‘like’, ‘get’, ‘think’ (removed)

Model Selection

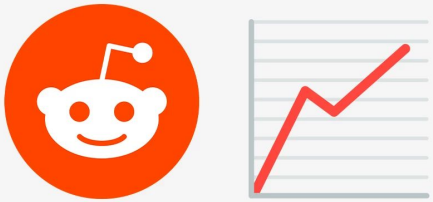
1. 4 optimised models were compared
    a. NMB with ngrams: {'cv__max_features': 32500, 'mnb__alpha': 1.4}
    b. TF-IDF LR: {'lr__C': 1.2, 'lr__penalty': 'l1', 'lr__tol': 0.35, 'tf__max_features': 25000}
    c. KNN
    d. Random forest
2. Models were scored using accuracy against an unseen test set
3. Selected model params with best estimators:
    a. NMB: {'cv__max_features': 32500, 'mnb__alpha': 1.4}

Key Findings

Based on the optimised TF-IDF Logistic Regression, we observed the following:

1. Previous generation owners seemed to be most interested in the newer version of their respective consoles: PS4 to PS5, Xbox One X to Xbox Series X
2. These game titles appeared to classify for the Xbox
    a. Destiny Series
    b. Doom Series
3. These game titles appeared to classify for the PS5
    a. Marvel's Spider-man Series
    b. Bloodborne
    c. Demon Souls Series

Conclusion



1. Top 3 key words for each subreddit
    a. XboxSeriesX subreddit: xbox, series, microsoft
    b. PS5 subreddit: ps5, ps4, playstation
2. Accuracy of 91% on test data where a given post came from
3. Trace posts with missing post labels back to their respective subreddit
4. Serves as a backup plan during a server glitch
5. Further improvements

The End

Q & A