Console Subreddits Classification
XboxOne vs PS4

Dominika Jones
Data Scientist
General Assembly

PROBLEM STATEMENT:

- Help lost souls in the market for a new gaming system and just joining this console generation.
- Create and train a classifier model on which subreddit a given post came from.

Subreddits

- Xbox One
- PS4 (PlayStation 4)

COLLECTING THE DATA

- Using api.pushshift.io
  - r/XboxOne
  - r/PS4

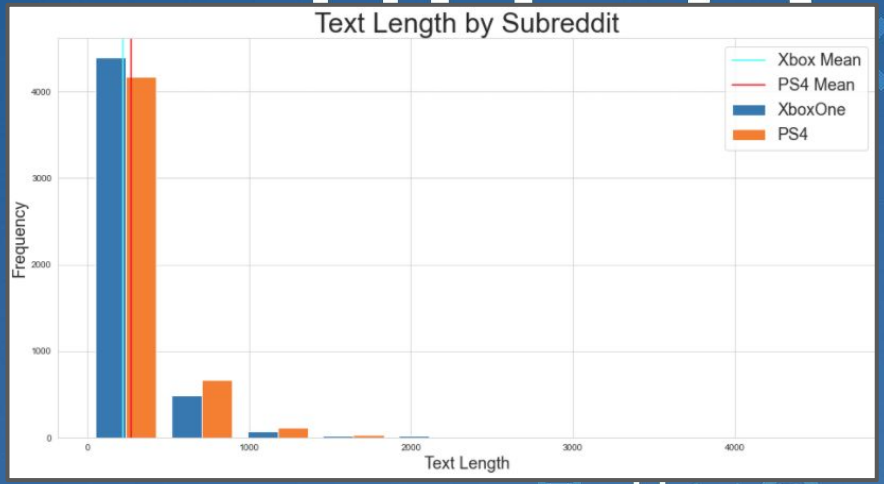
- Collected 10000 posts
  - 5000 each
  - Title, selftext, id

SUBREDDIT POSTING RULES

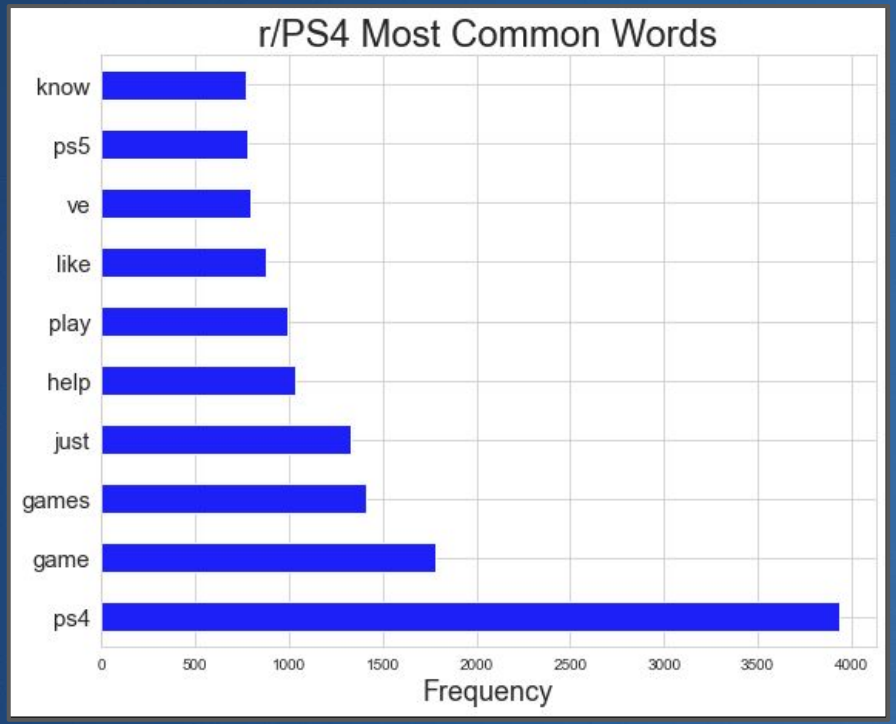
	R/PS4 RULES	R/XBOXONE RULES
1	Do not personally attack other users.	Posts must be related to Xbox and link directly to the source.
2	Do not discuss hacking, piracy or grey market.	Follow Reddiquette. Keep it civil. Spoilers and NSFW must be properly marked.
3	All spoilers and NSW context must be marked.	Piracy, hacking, jailbreaking etc.
4	Do not submit questions with vague titles.	No memes, image macros, polls, petitions, friend requests, etc.
5	Media submissions must use proper formatting.	No advertising, selling, buying, trading, self promotion, or asking for free stuff.
6	All suggestions require means to cause change.	[Giveaway] and [Deal] posts
7	Do not submit low effort content.	[Tech] posts must be within the weekly [Tech] megathread
8	Review Threat and Giveaway Requirements.	Showoff Sunday
9	Event and Megathread Policies.	
10	Don't post spam	
11	No PS5 discussion.	

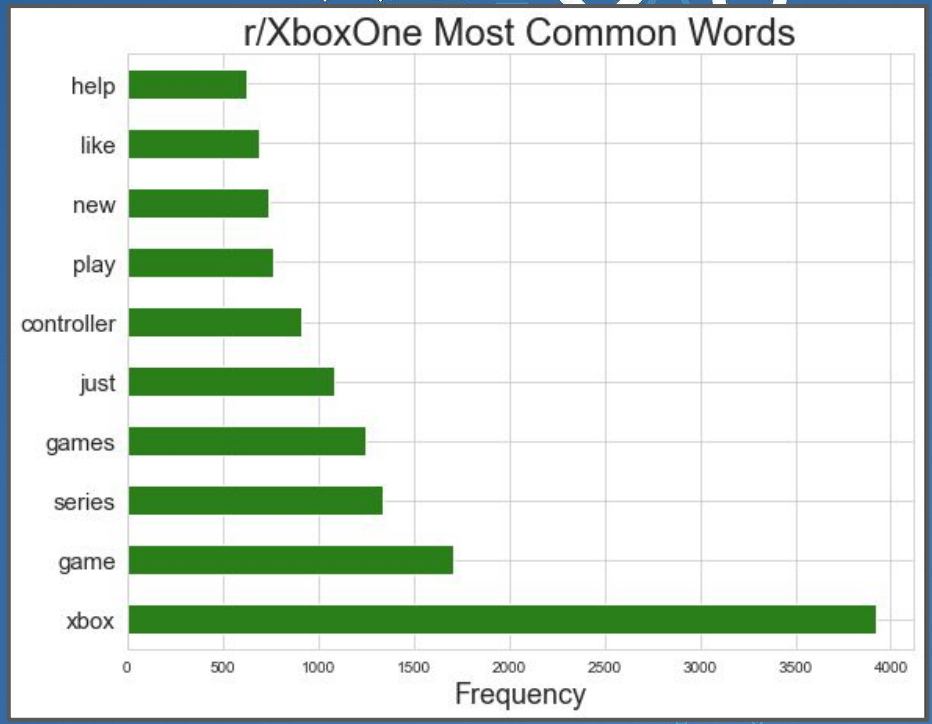
DATA CLEANING/ PREPROCESSING

- Merged title and self text to deal with null and [removed]
- Binarized subreddits
- Applied:
  - Regular Expression
  - Removing Stop words
  - Lemmatizing

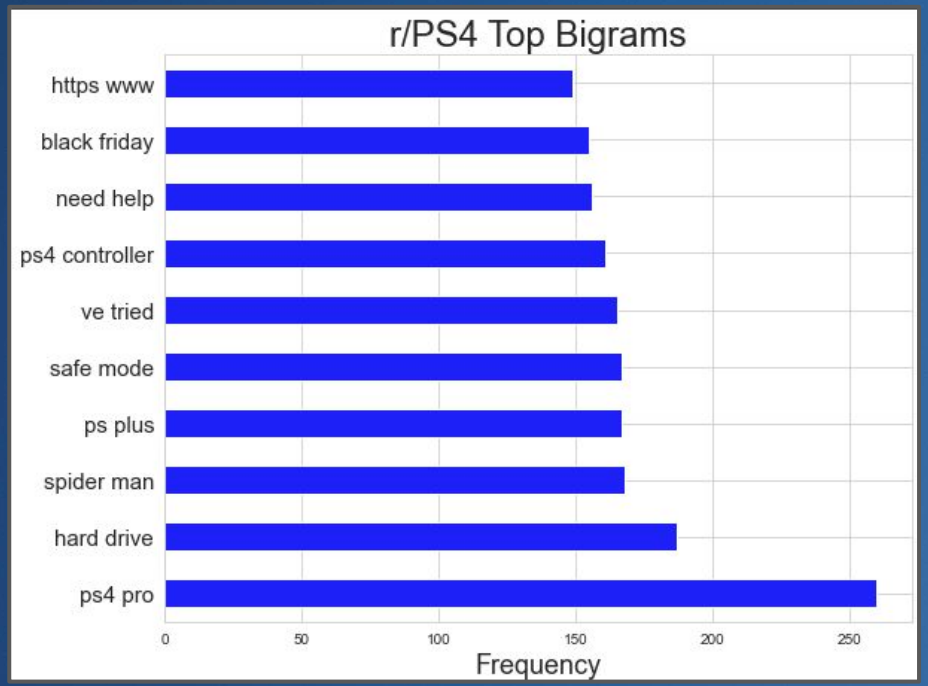


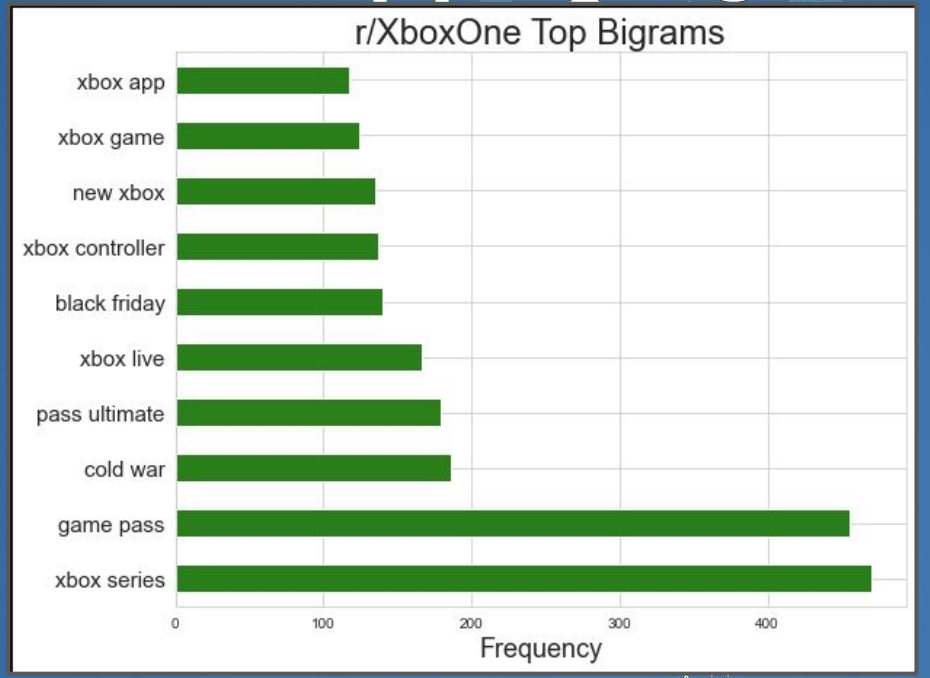
10 MOST COMMON WORDS IN EACH SUBREDDIT



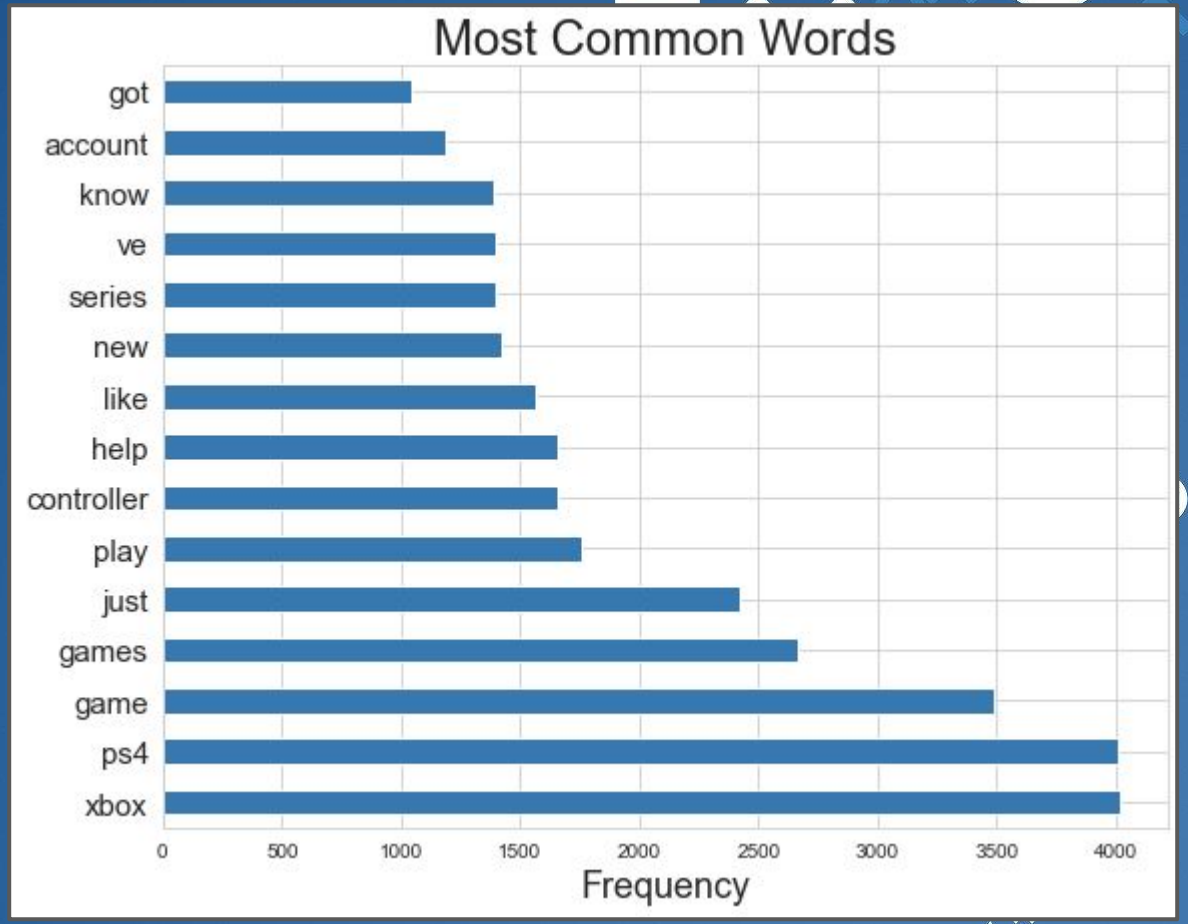


10 MOST COMMON BIGRAMS IN EACH SUBREDDIT



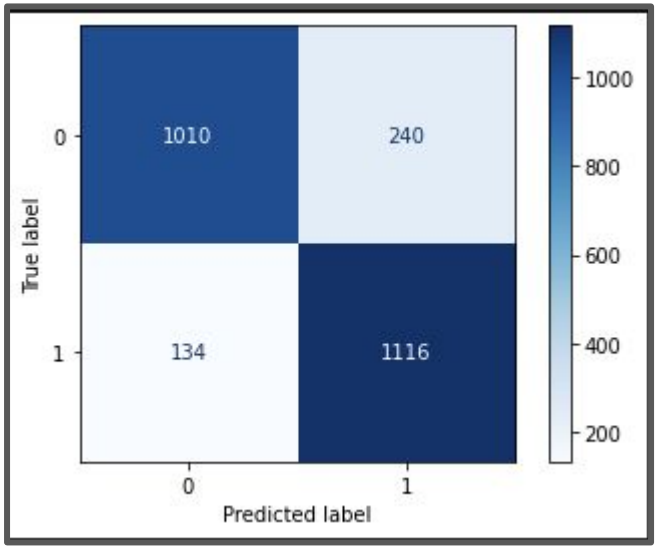


Most commonly occuring words in the dataframe



MODELS

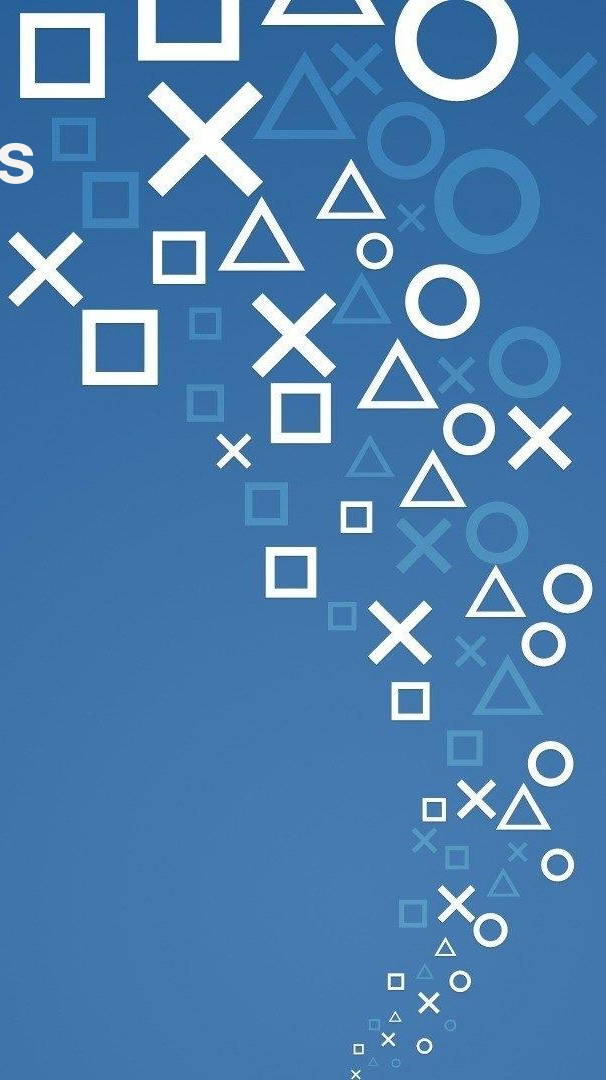
- Baseline accuracy
  - 0.5
- Best Model
  - Naive Bayes & CountVectorizer
  - With a 0.85 score



MODEL	TRAIN SCORE	TEST SCORE	SPECIFICITY	SENSITIVITY
Logistic Regression	0.985	0.868	0.893	0.842
Naive Bayes	0.966	0.85	0.808	0.89
Random Forest	0.991	0.848	0.858	0.838

Conclusion & Recommendations

- Naive Bayes Model performed the best and should predict a post subreddit with high accuracy.
- Explore parameters further
- Expand stop words list
- Look into misclassifications
- Sentiment analysis



References:

- Xbox One vs PS4 vs Switch: Console and Game Sales Numbers – 2020
- PlayStation 4 - News • Discussion • Community
- /r/XboxOne - The home of everything Xbox One on Reddit