Project 3: Web APIs and NLP

Lee Mei
Tze Yi
Mei Lian

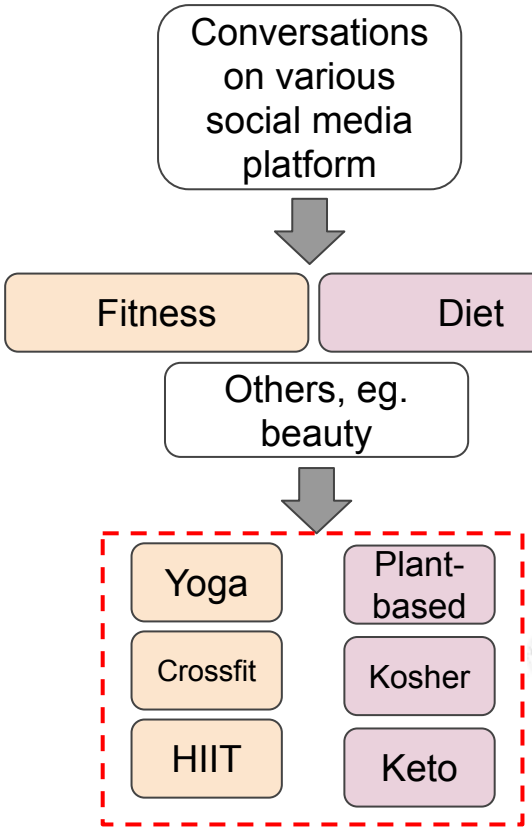
Problem Statement & Objective

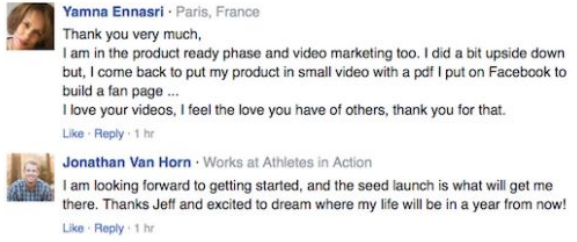
Problem statement:

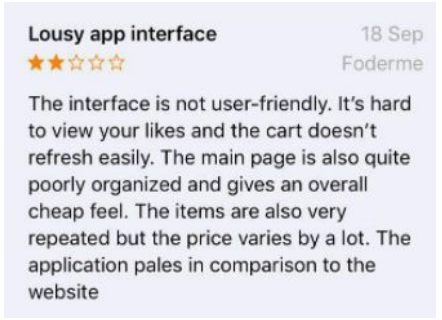
- Classifying various topics of discussion on different social media platforms into their own sub-categories

Objective:

- Develop a classification model which is able to classify/categorise subtopics based on the netizens' comments / discussion
- The model should be able to distinguish / classify discussion/comments on topics which are closely related
- The model which can produce the best accuracy will be shortlisted and trained further to include more subtopics







Overview

Goals/Objectives

Develop a classification model which can categorise data into sub-categories

Data Scraping

Data selected → r/Netflix vs r/AmazonPrimeVideo from Reddit

API push shifts to scrape the data

Data Processing

Process the scraped data:

- Missing values
- Cleaning
- Tokenisation
- Stop words
- Lemmatise/ Stemming

Data Exploration & Modeling

Models evaluated:

- Logistic Regression
- Random Forest
- Multinomial Naive Bayes

Conclusion

The best performing model is selected for deployment

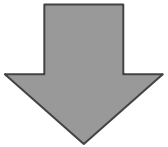
Further training on the model to classify more sub-groups

Data scraping - r/Netflix

https://www.reddit.com/r/netflix/

941k members

Created since 21 Nov 2008



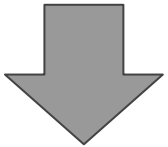
	subreddit	selftext	title
0	netflix	[removed]	Any movie or series recommendations
1	netflix	I want to download from Netflix, but not by ju...	Downloading
2	netflix	[removed]	Downloading
3	netflix		Season 5 of Kim's Convenience is now on Netflix.
4	netflix	I've used other people's Netflix accounts for ...	How do the Netflix subscription accounts work?

Data scraping - r/AmazonPrimeVideo

https://www.reddit.com/r/AmazonPrimeVideo/

3.3k members

Created since 20 Aug 2013



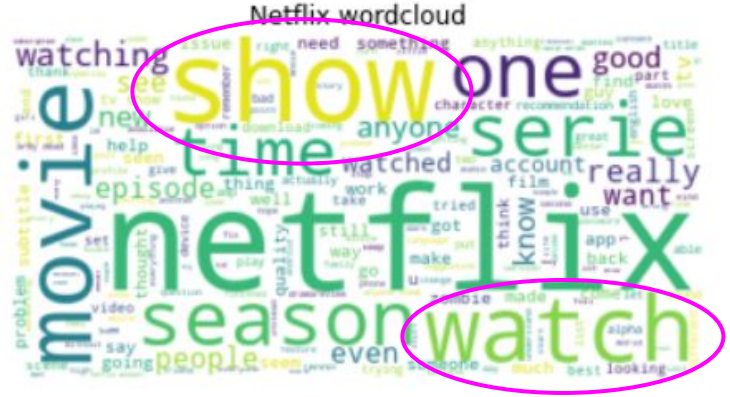
	subreddit	selftext	title
0	AmazonPrimeVideo	[removed]	Different video selections on different devices
1	AmazonPrimeVideo	I can't find the English sub or dub version fo...	Can't find the English version for Naruto (Can...
2	AmazonPrimeVideo		Mary J Blige's My Life - Official Trailer | Pr...
3	AmazonPrimeVideo		Producers on Amazon's 'Prisma' Where Twins Cha...
4	AmazonPrimeVideo	New to prime video im watching “the office”. I...	Prime Video is skipping episodes and seasons!

Data processing

- Missing / redundant values
- Symbols / numbers
- Lowercase
- Punctuations
- Stopwords
- Lemmatise and checking whether stemming is required

	title	selftext	subreddit
0	Any Stack TV subscribers here?	[removed]	AmazonPrimeVideo
1	chromecast	[removed]	AmazonPrimeVideo
2	Using digital reward credit for channel subscr...	I've got some digital rewards for no rush ship...	AmazonPrimeVideo
3	Infinite loading during ads.	[removed]	AmazonPrimeVideo
4	Infinite loading screen whenever an ad tries t...	NaN	AmazonPrimeVideo
5	Anyone Know What Time Clarkson's Farm On Prime?	Hi anyone know what time (UK or EST) Clarkson'...	AmazonPrimeVideo
6	Confused	So, I'm a bit confused here. I am watching YuG...	AmazonPrimeVideo
7	So, channel subscriptions do not include all t...	I've been rewatching the Mythbusters series an...	AmazonPrimeVideo
8	Why I Have Mixed Feelings Over The Family Man ...		AmazonPrimeVideo
9	The creepiest movie from my childhood is on Pr...	It's called The Adventures of Mark Twain. It's...	AmazonPrimeVideo

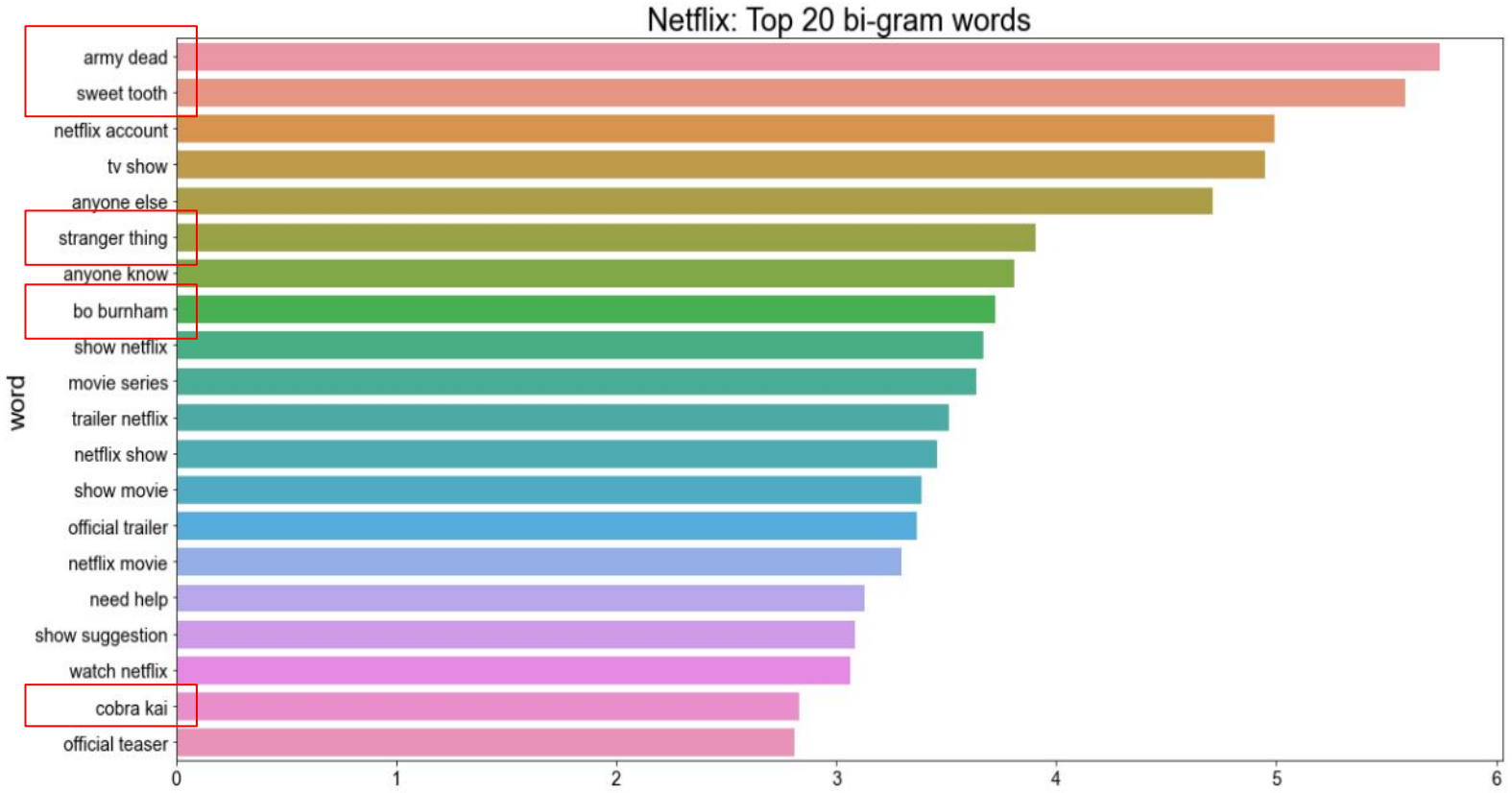
Data Exploration - WordCloud



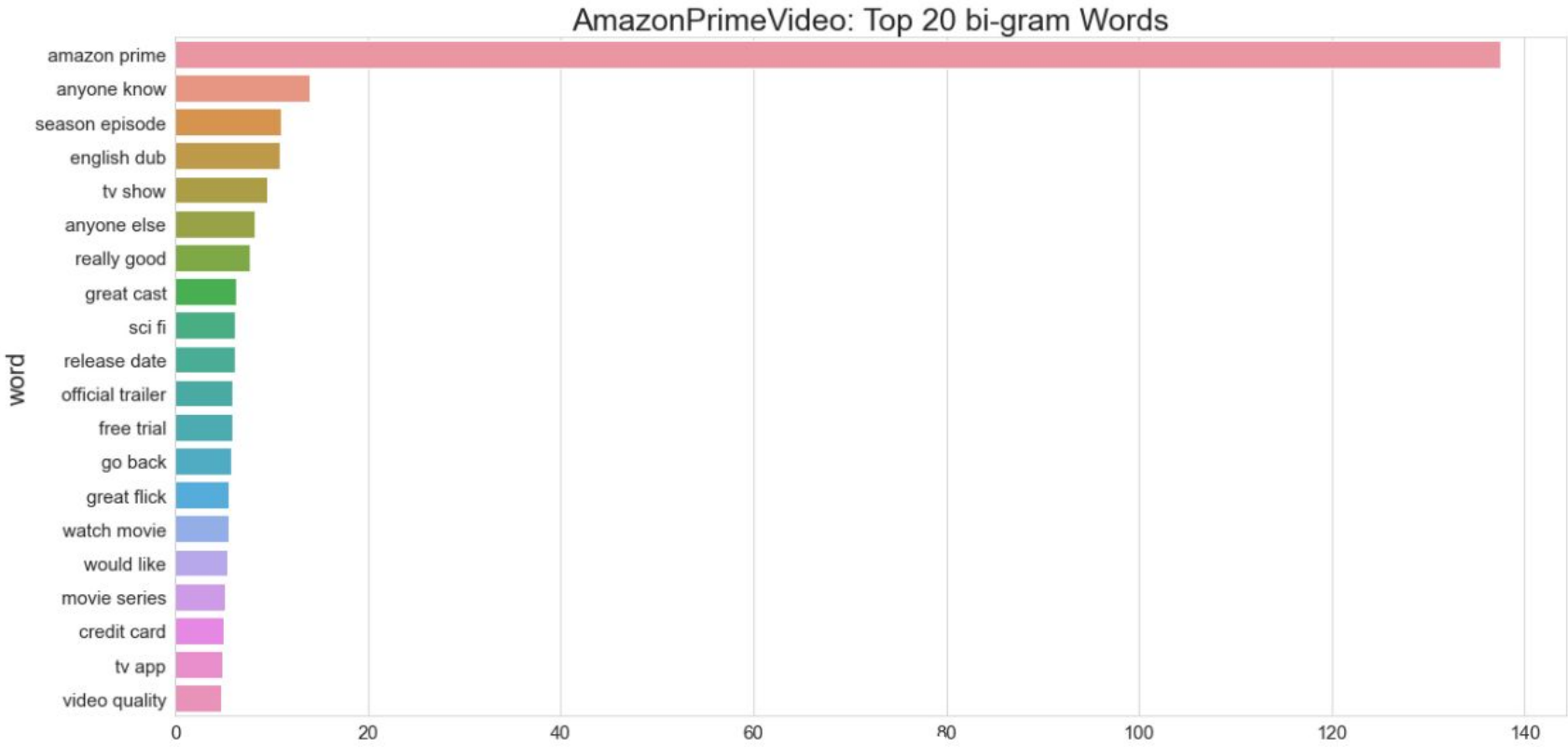
Distinct words such as ‘netflix’, ‘prime’, ‘amazon’

Seems like both subreddits are equally popular in terms of movies and TV shows

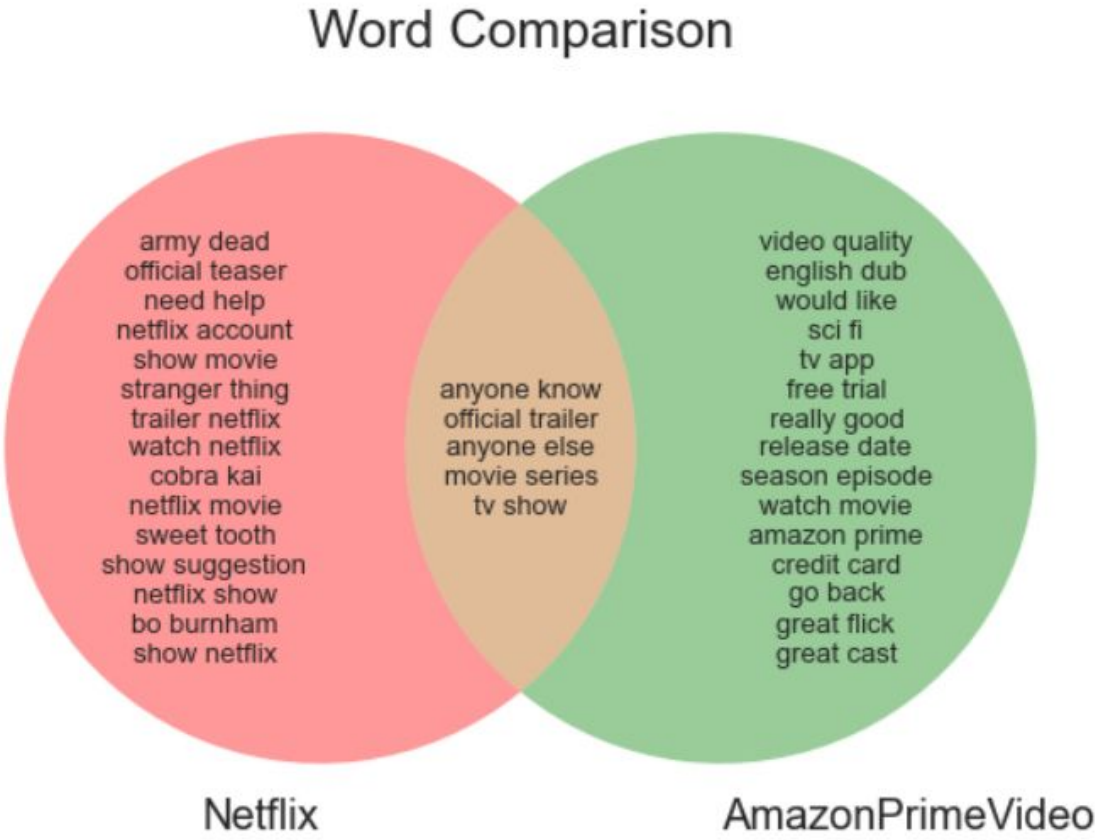
Data Exploration



Data Exploration



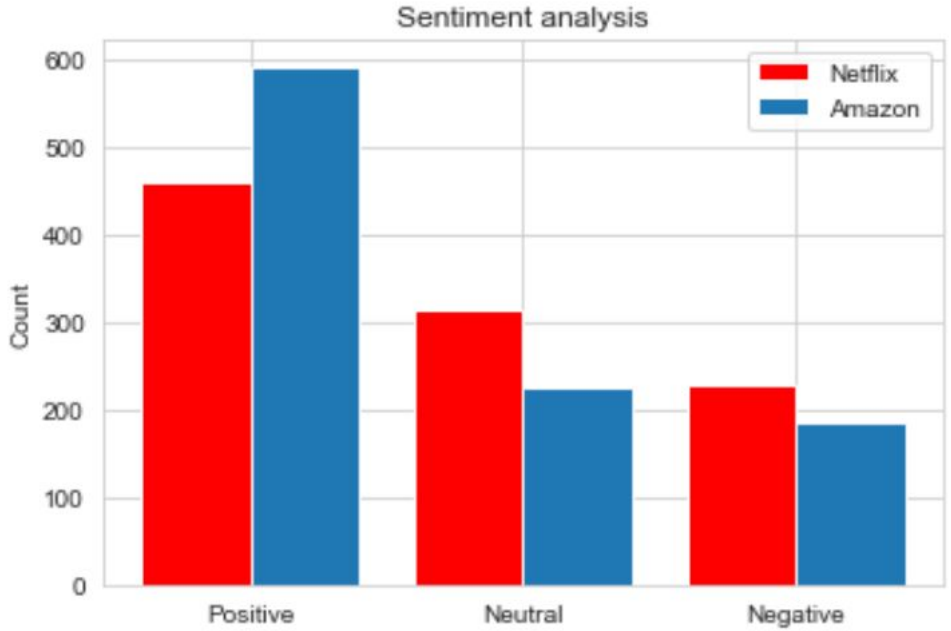
Data Exploration - Netflix vs Amazon Prime Video



Overall, both channels are popular for their movies and tv shows.

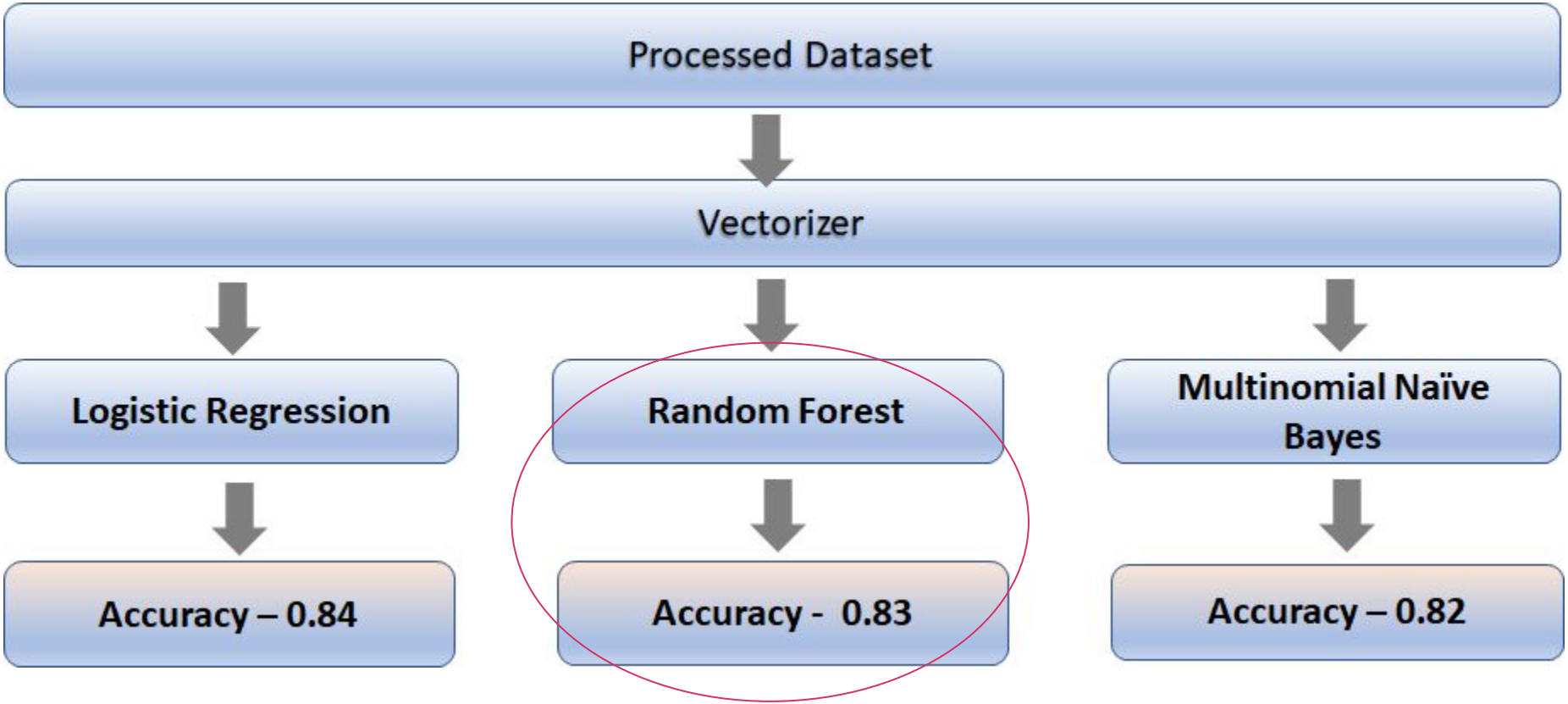
However, Netflix seemed to be known for their original tv shows.

Data exploration - Sentiment analysis



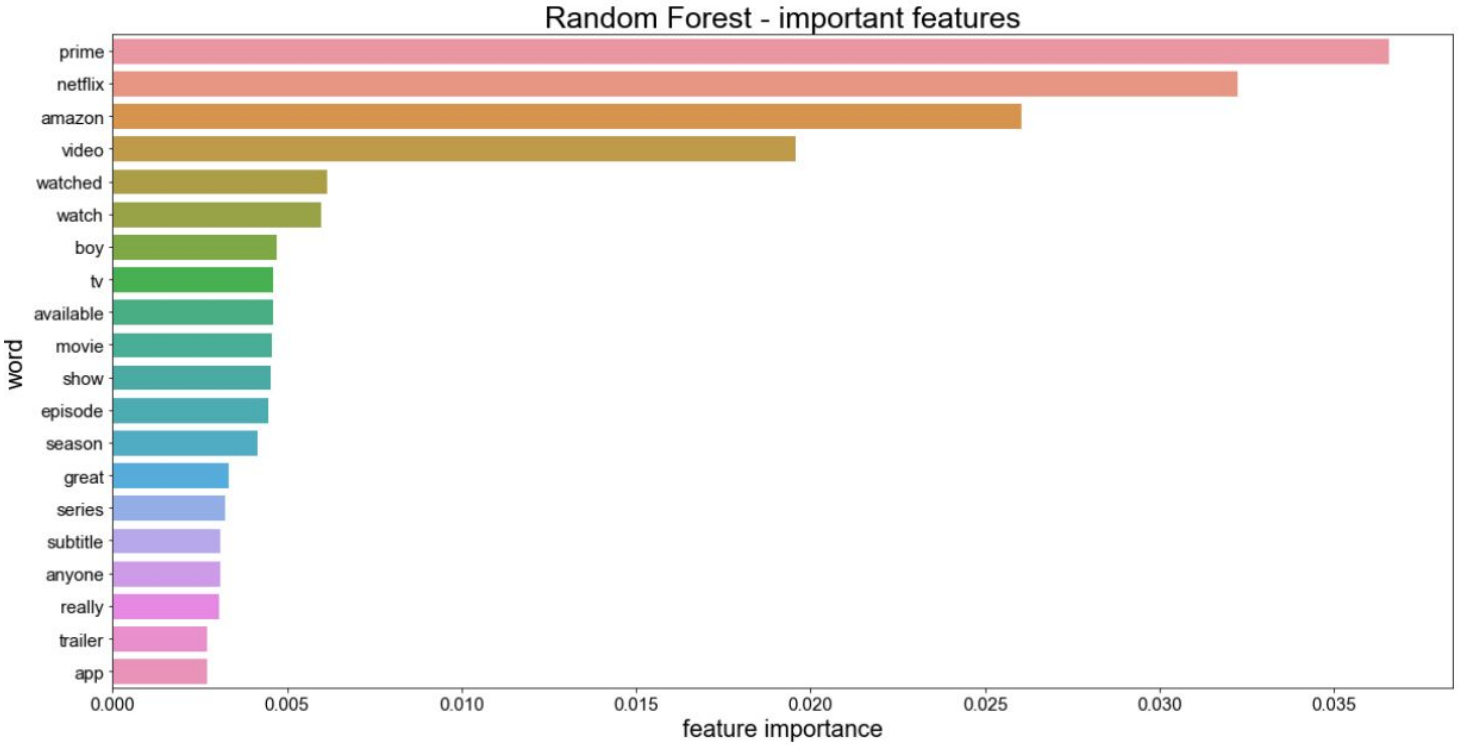
Overall, there seem to be more positive sentiment for Amazon Prime Video compared to Netflix. Its negative sentiment is also lower than Netflix

Modeling



✓ Dynamic → can classify more categories
✓ Can map non-linear relationship

Modeling - Top features under Random Forest



Interpretation:

Feature importance provide us an indication of how important a word is when it comes to **splitting** the trees in the random forest but the feature importance does not indicate the likelihood of which category

Conclusion

Observations

- Both streaming services are popular with movies and tv shows
- Netflix has branded itself with its offering of original shows

Recommendation

- Deploy the Random Forest model
  - more dynamic, i.e. it can classify more categories
  - can map non-linear relationships

Going forward

- Train the model with more data so that it can classify more subtopics

Limitation

- Requires frequent training on hot topics / latest trends (hence costly to maintain)

Q&A

Thank you