

The background of the slide is a composite image. On the left, there is a close-up of a film reel with its sprocket holes visible. On the right, there is a clapperboard with the words 'PRODUCTION', 'DIRECTOR', 'CAMERA', 'SCENE', and 'TAKE' printed on it. A semi-transparent white rectangular box is centered over the image, containing the title and author's name.

Movie&TV Reviews Topic Modeling

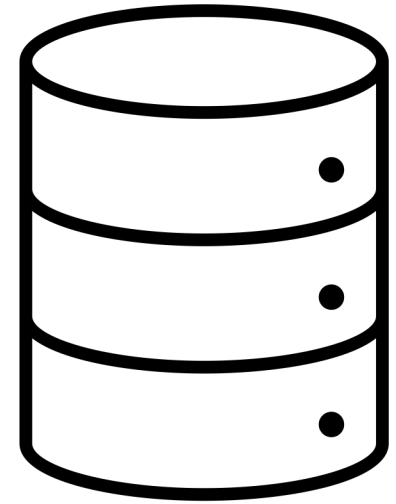
Rui Yuan

Clients and Problem

- Amazon wants to investigate customer reviews and improve products
- The goal is to find out the overall customer sentiments and topics customer care about.

Data

- Data accessed through
- 16905 rows of customer reviews
- Features: customer ID, rating (1-5), review text



Work Flow and Algorithms

EDA and Data Visualization:

- Matplotlib, WordCloud

Text Preprocessing

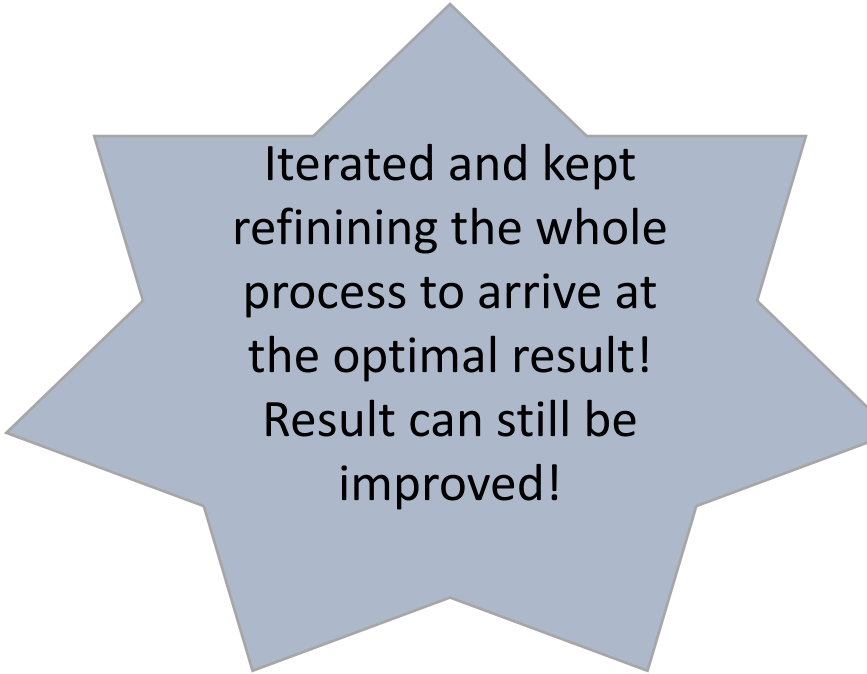
- NLTK used including word tokenization, lemmatization, pos tagging, stop words removal
- Filtered words down to nouns, adjectives, verbs and adverbs

Sentiment Analysis:

- VADER, TextBlob used and settled with VADER

Topic Modeling:

- With only nouns
- LDA, LSA, NMF, CorEx used and settled with NMF



Iterated and kept refining the whole process to arrive at the optimal result! Result can still be improved!

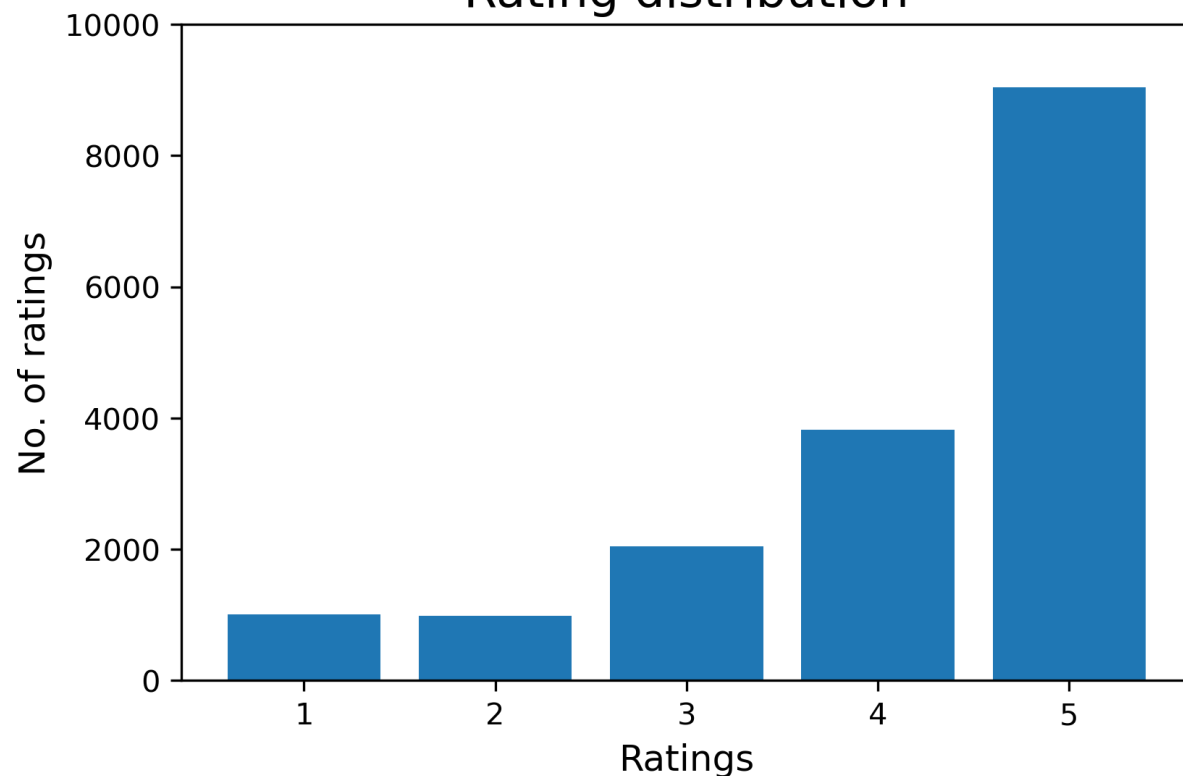
Common Words Visualization: WordCloud



- Positives more than negatives
- Movie elements: story, series, character, action, music, plot, role, performance, scene...

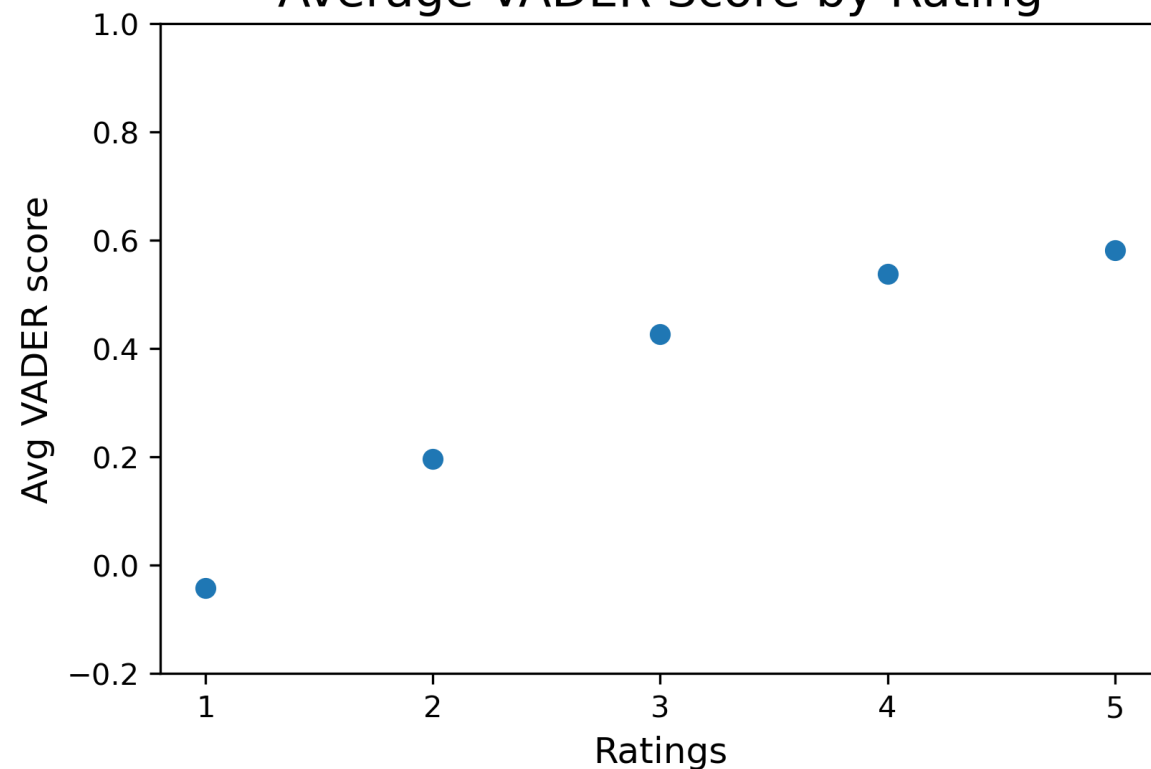
☹️ Review Sentiment: VADER

Rating distribution



- Many high ratings
- Customers are relatively happy with the products

Average VADER Score by Rating



- Higher ratings, higher VADER score
- VADER seems to correctly detect the sentiments

Topic Modeling: NMF

Future Work

- Continue refining topic modeling and exploring more NLP methods and tools such as Spacy, ScatterText, etc.
- Build Content and Collaborative Recommender System