IMDb watchlist web scraping

Goal to achieve

- Collect every item on my personal Imdb watchlist
- Put every relevant information in a dataframe
    - Title, Rating, Release year, Plot, Summary
- Handle missing items
- With the help of AI - do a sentiment analysis based on the short Plot and on the Summary
    - Sentiment_analyzer
    - ChatGPT - prompt
- Compare the results of the analysis

Data + difficulties

- Website - Watchlist
- The basic list contains 2723 items → **2491** items were used for the analysis
    - 221 items were removed during the cleaning process due to missing year
    - 5 items were removed during the cleaning process due to missing plot (2 were removed in the previous step)
    - 12 items were removed during the cleaning process(3 were removed in the previous steps)

- The base of the data is my personal watchlist on IMDb
- The website alters the backend code from time to time that requires code modification on a regular basis
- I started the project in Google colab, however due to usage restrictions it had to be rebuilt in jupyter notebook
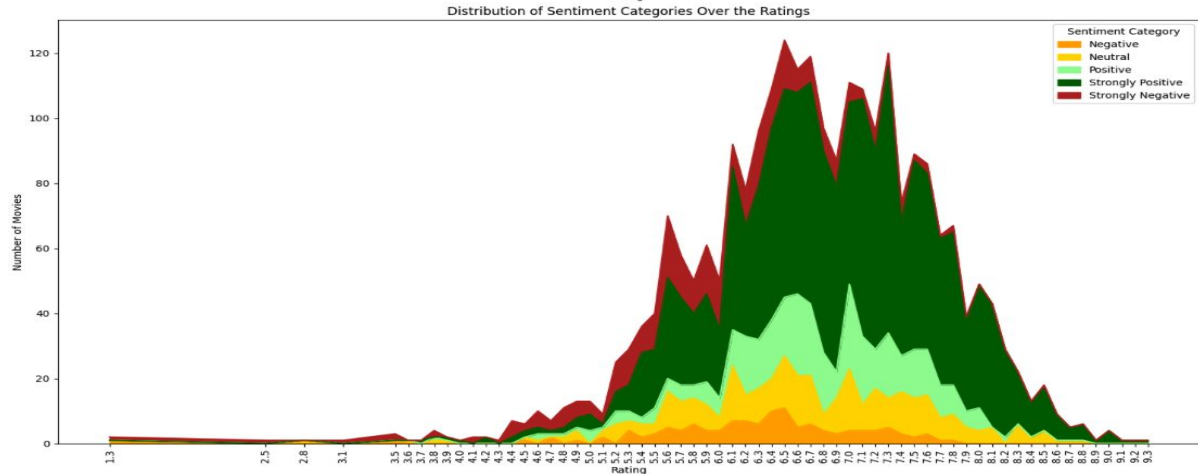- Sentiment analysis - AI part of the project provides ambiguous results

Results / 1

Plot-based Sentiment Categories

- Strongly Negative → **1,056 movies**
- Negative → **30 movies**
- Neutral → *(not detected)*
- Positive → **39 movies**
- Strongly Positive → **1,366 movies**
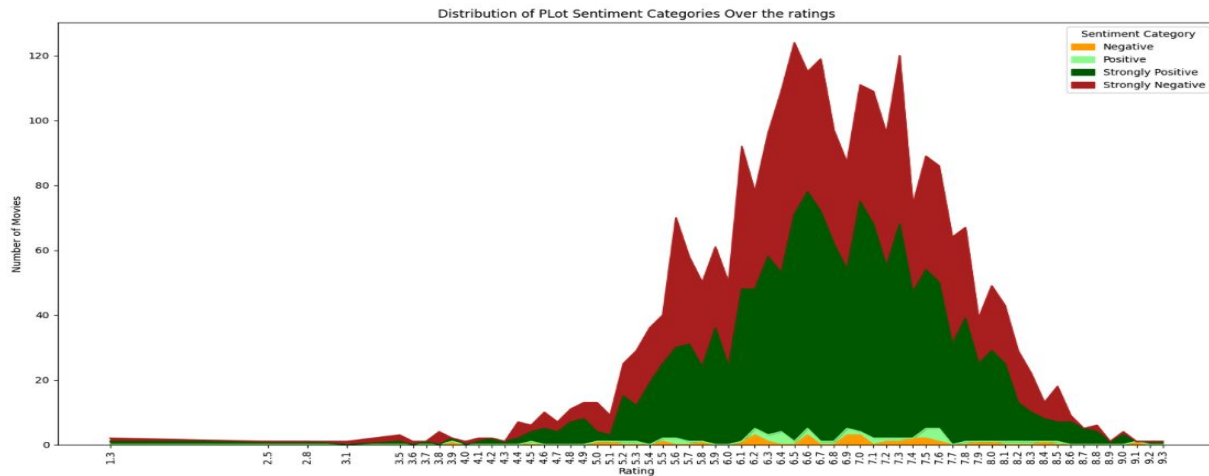
Summary-based Sentiment Categories

- Strongly Negative → **281 movies**
- Negative → **126 movies**
- Neutral → **308 movies**
- Positive → **357 movies**
- Strongly Positive → **1,419 movies**

Distribution of PLot Sentiment Categories Over the ratings

Distribution of Sentiment Categories Over the Ratings

Results / 2

Chat GPT results Emotional categories

- Strongly Negative → **448 movies**
- Negative → **861 movies**
- Neutral → **461 movies**
- Positive → **721 movies**

Distribution of Sentiment Categories based on the Ratings
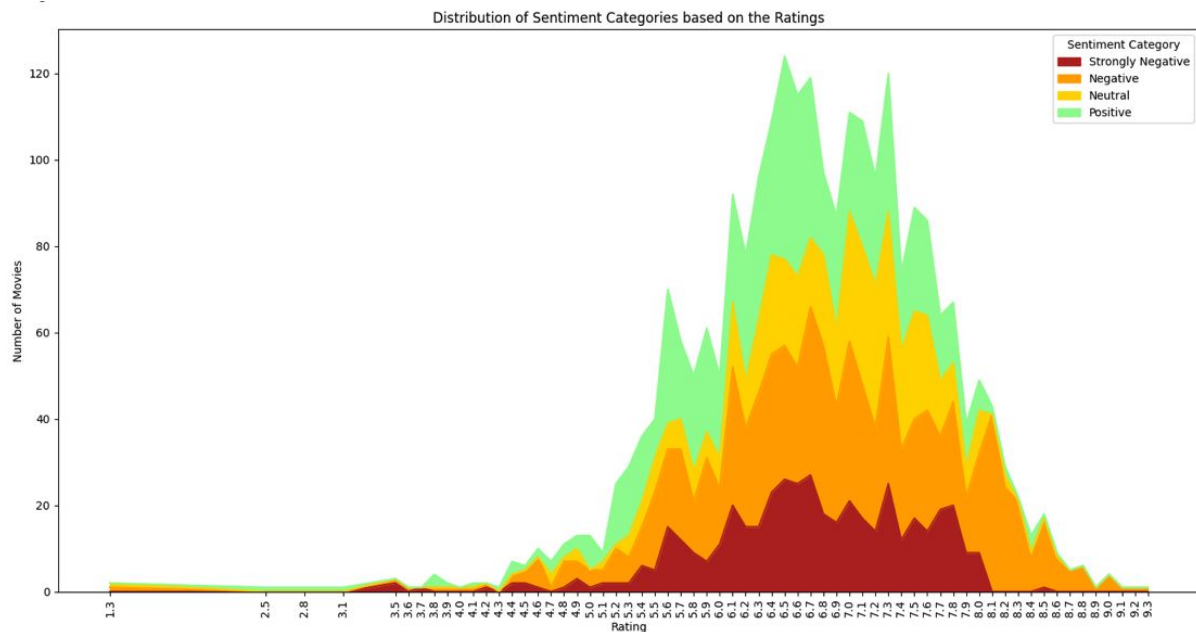
Summary

- ChatGPT was working with the Plot only

- According to the sentiment analyzer
  - 43.59% of the movies are Negative
  - 56.41% of the movies are Positive
- According to ChatGPT the distribution is more even
  - 52.54% of the movies are Negative
  - 28.94% of the movies are Positive

🎯 SENTIMENT ANALYSIS RESULTS
===

📈 PERCENTAGE DISTRIBUTION
----

sentiment_category	plot_percentage	summary_percentage	gpt_percentage
Strongly Negative	42.39%	11.28%	17.98%
Negative	1.2%	5.06%	34.56%
Neutral	0.0%	12.36%	18.51%
Positive	1.57%	14.33%	28.94%
Strongly Positive	54.84%	56.97%	0.0%

📋 RAW COUNT DISTRIBUTION
----

sentiment_category	plot_count	summary_count	gpt_count
Strongly Negative	1056	281	448
Negative	30	126	861
Neutral	0	308	461
Positive	39	357	721
Strongly Positive	1366	1419	0

📊 SUMMARY STATISTICS

📝 Total Summary Records: 2,491
📖 Total Plot Records: 2,491
🎂 Total GPT Records: 2,491

🏆 Most Common Sentiments:
Summary: Strongly Positive
Plot: Strongly Positive
GPT: Negative

Lessons learned / next steps

- Reduce duplicates during web scraping
- Handle the missing year items (a series has a start and end date)
- Other alternatives to sentiment analysis → chatGPT / claude
- Re-run the analysis multiple times to compare the results
- Revise the scraping code to make it more efficient
- Use the IMDb API for more efficiency