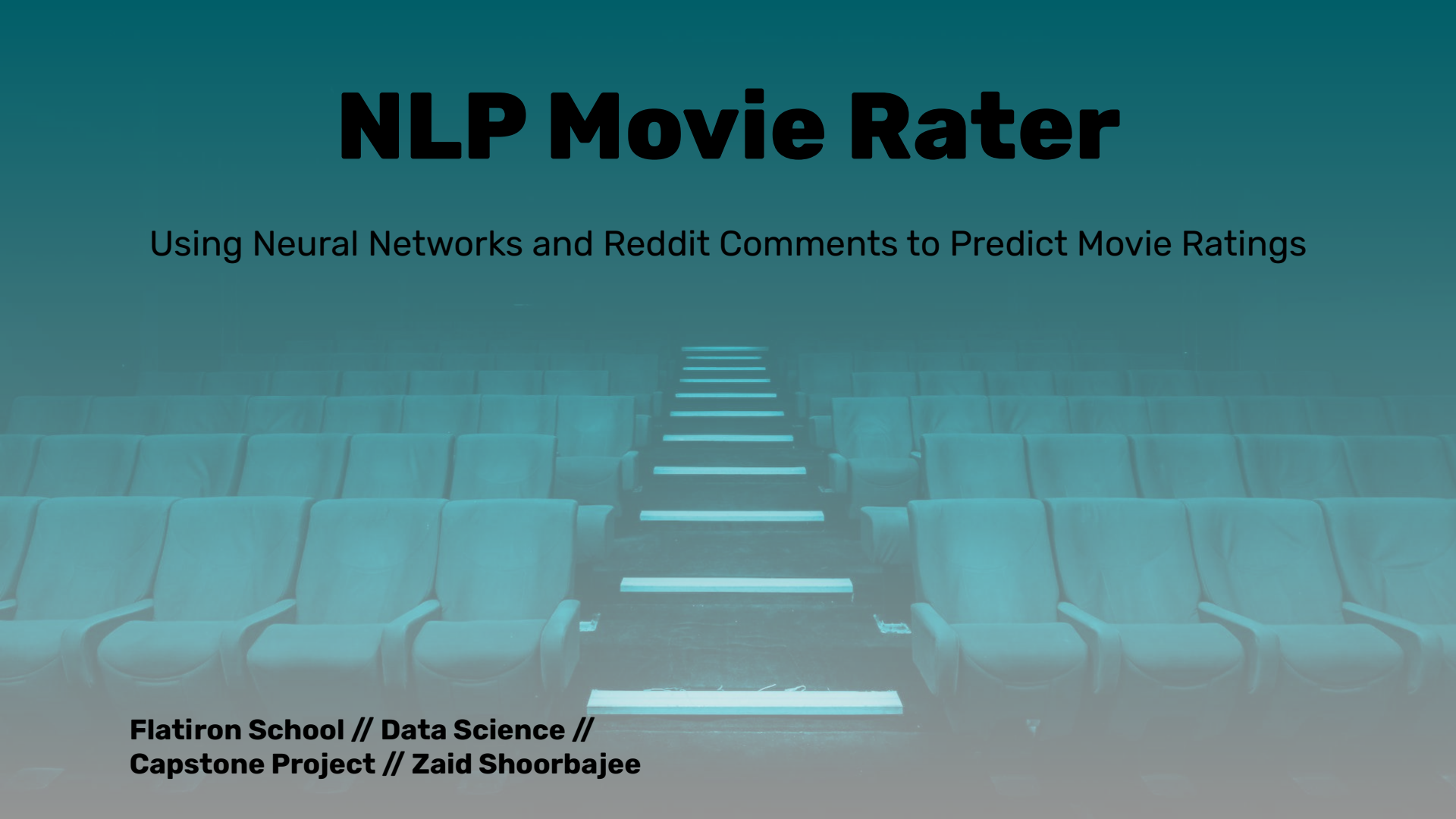
NLP Movie Rater

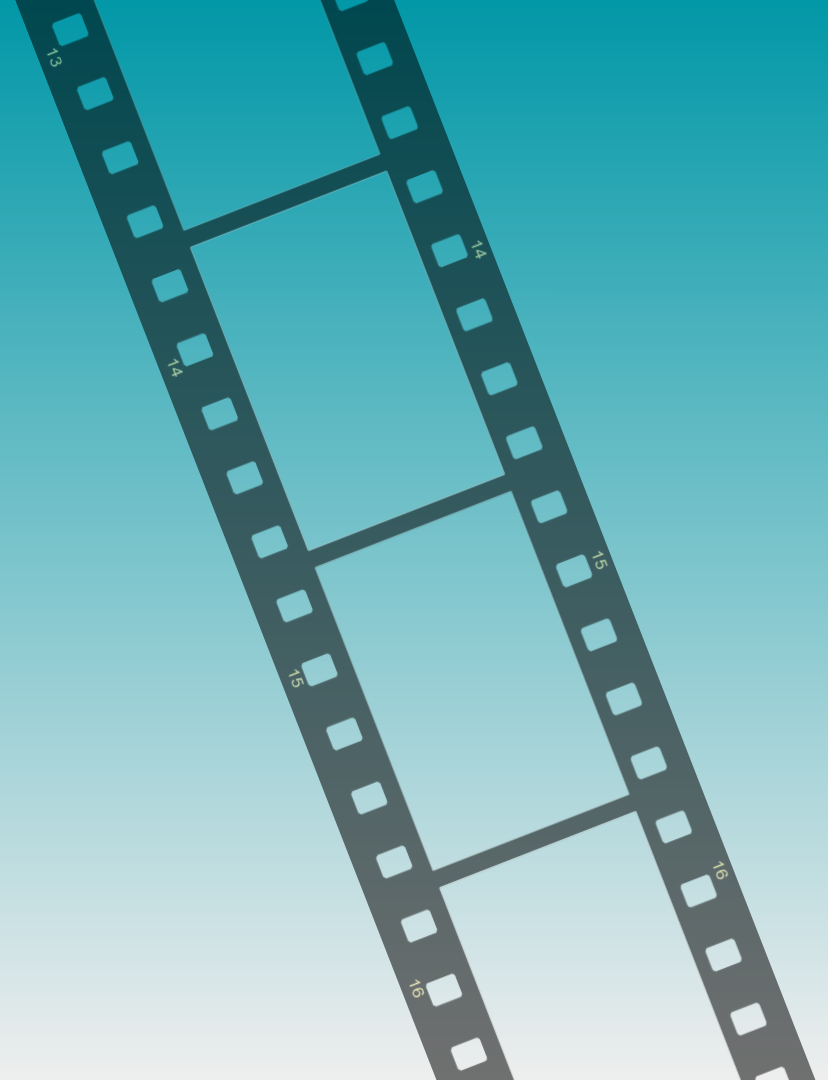
Using Neural Networks and Reddit Comments to Predict Movie Ratings



Flatiron School // Data Science //
Capstone Project // Zaid Shoorbajee

Overview

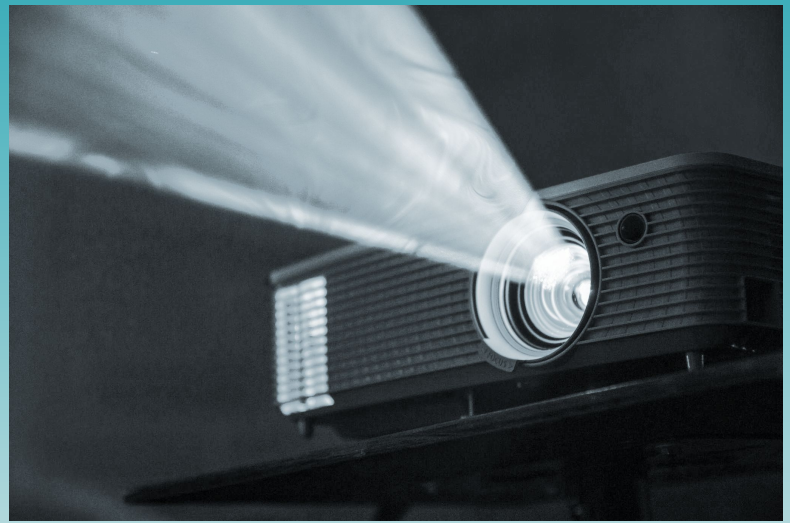
- Business Understanding
- Data Understanding
  - Dataset
  - NLP
- Modeling
- Evaluation
- Recommendations
- Limitations & Future Work



Business Understanding

A movie studio, *Cinedistance*, wants to know the critical reception of its movies before they are released.

- Plans to release movies early to a focus group of 100 people.
- Prompt: What would you say about this movie in an internet comment section?
- Data Science task: Predict the movie's IMDb score based on the focus group's feedback.



Business Understanding

Why predict movie ratings?

This can inform decisions around marketing, distribution, re-shoots, edits, etc.

Data Understanding: Dataset

Reddit's **r/movies** community

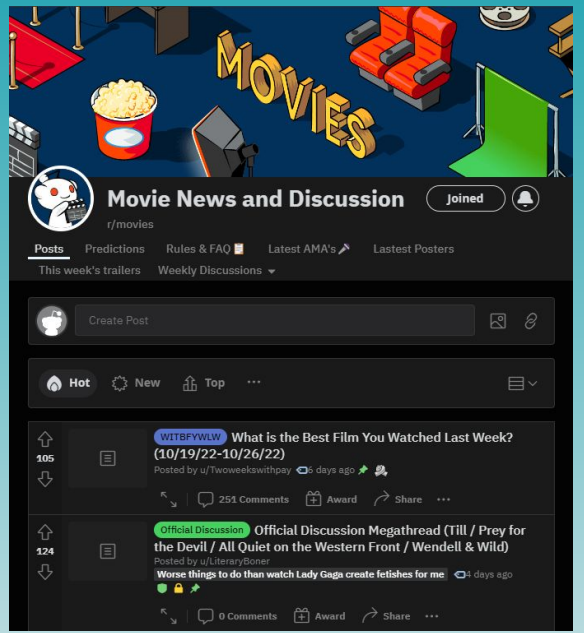
- ~30 million members
- Official discussions of most major releases
- Scraped top 100 comments from as many movie discussions as possible

Features:

- 70,693 text comments from 922 movies
- Years: 2016-2022

Target:

- IMDb movie ratings (1-10)
- Downloaded from IMDb website
- Mean rating: 6.47



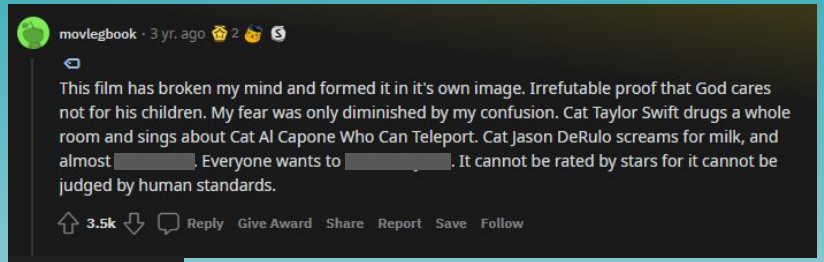
IMDb

Data Understanding: NLP

Natural Language Processing

- Language is messy
- Easy for humans to understand; not for machines
- Unstructured data
- Must be heavily processed

Example r/movies comments



Cats (2019)

The Batman (2022)

Data Understanding: NLP

🌟 Original Comment 🌟

The whole scene with The Riddler talking about “Bruce Wayne” had me hooked. I hope Paul Dano returns for the sequel.

Sentiment

POSITIVE

Tokenized

['the', 'whole', 'scene', 'with', 'the', 'riddler', 'talking', 'about', 'bruce', 'wayne', 'had', 'me', 'hooked', 'i', 'hope', 'paul', 'dano', 'returns', 'for', 'the', 'sequel']

Part-of-Speech

['DET', 'ADJ', 'NOUN', 'ADP', 'DET', 'NOUN', 'VERB', 'ADP', 'PROPN', 'NOUN', 'VERB', 'PRON', 'VERB', 'PRON', 'VERB', 'PROPN', 'PROPN', 'NOUN', 'ADP', 'DET', 'NOUN']

Lemmatized

['the', 'whole', 'scene', 'with', 'the', 'riddler', 'talk', 'about', 'bruce', 'wayne', 'have', 'i', 'hook', 'i', 'hope', 'paul', 'dano', 'return', 'for', 'the', 'sequel']

Lemmatized & Removed Stop Words

['scene', 'riddler', 'talk', 'bruce', 'wayne', 'hook', 'hope', 'paul', 'dano', 'return', 'sequel']

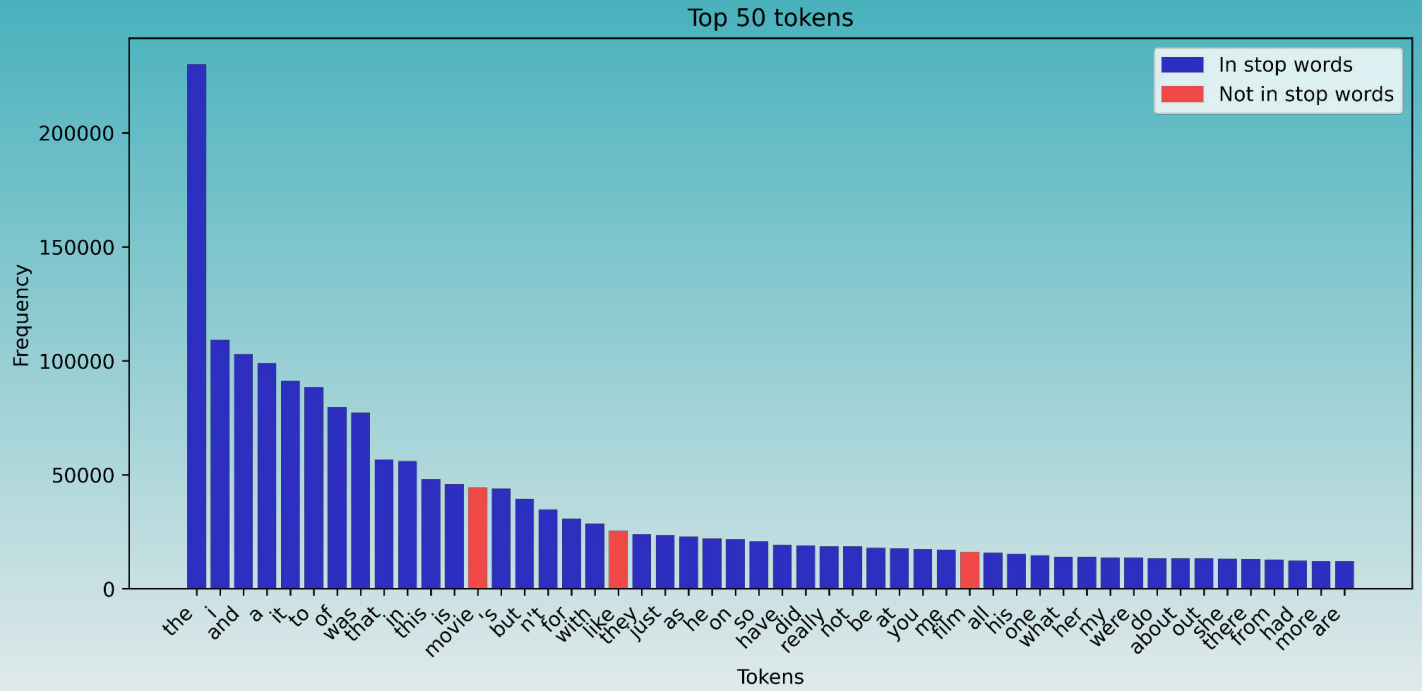
TF-IDF Vectorized (all comments)

	The Batman (TF-IDF)
city	0.197992
dark	0.180793
end	0.150690
feel	0.279118
know	0.122270
love	0.216030
people	0.123268
scene	0.275356
think	0.231655
way	0.124107

Plus, values of “0” for all other words in the corpus

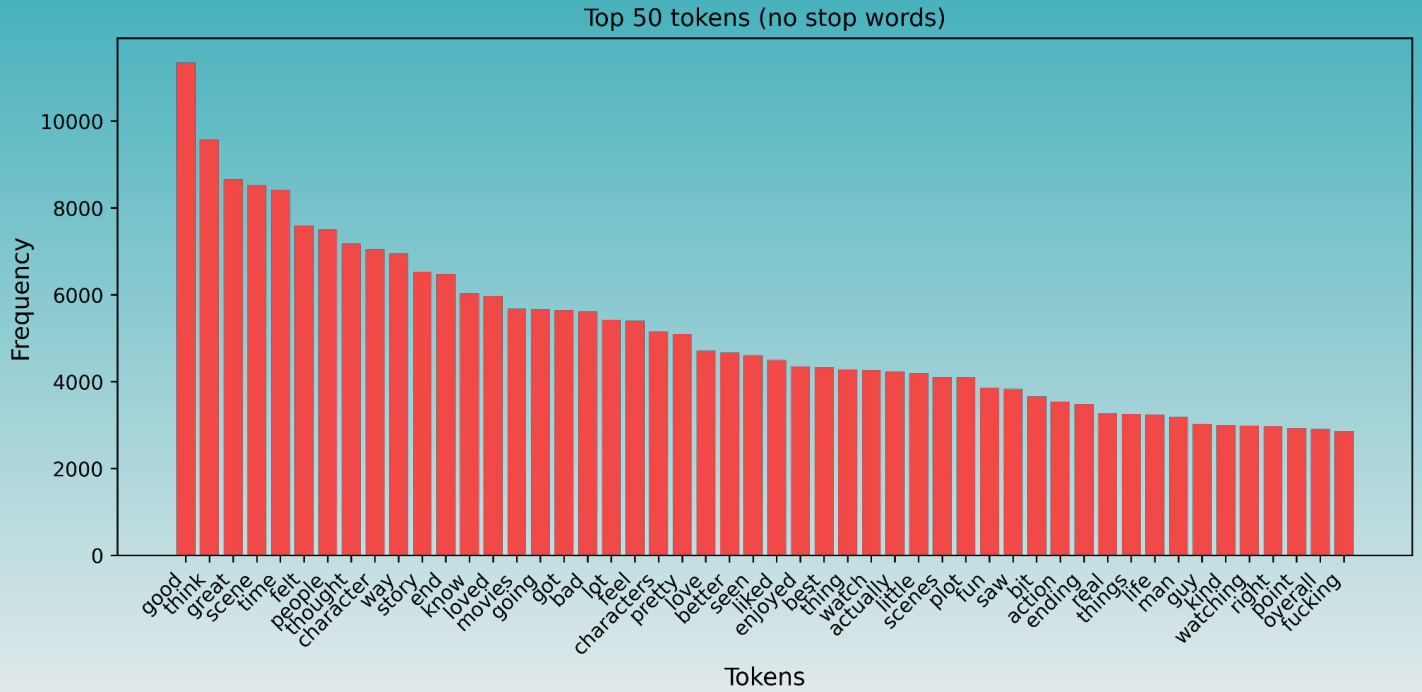
Data Understanding: NLP

Before removing stop words



Data Understanding: NLP

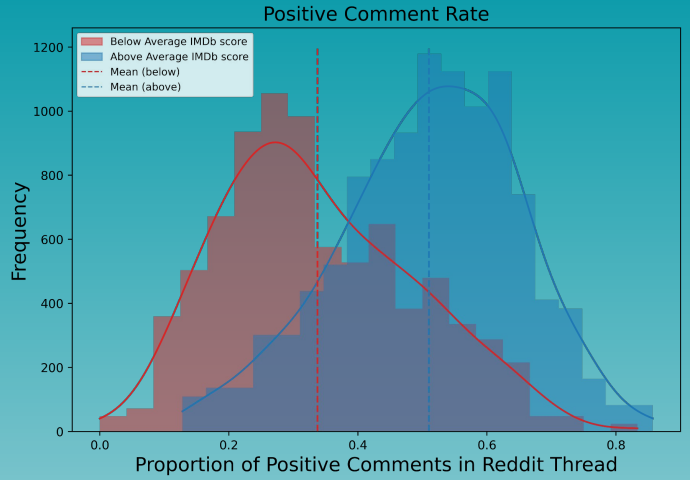
After removing stop words

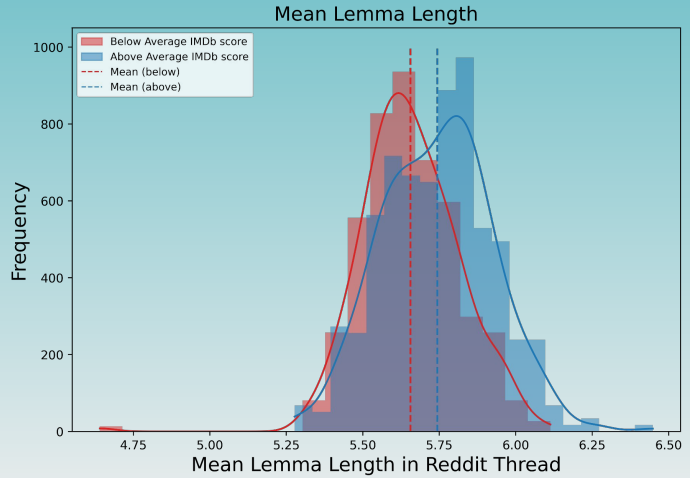


Data Understanding: NLP

Other features
(per comment section)

- Proportion of positive comments
- Proportion of negative comments
- Average character count of comments
- Mean length of lemmas
- Proportion of unique lemmas
- Proportion of stop words



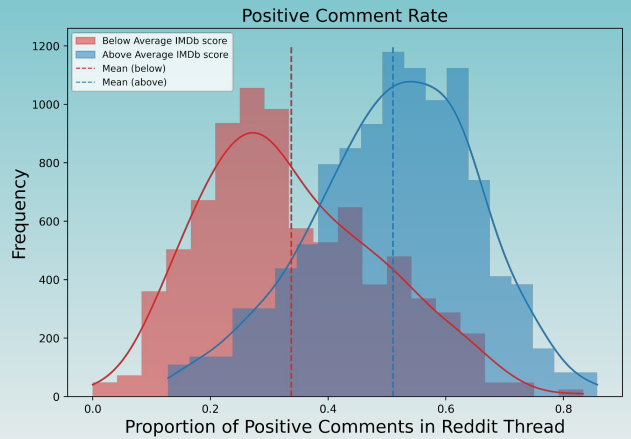


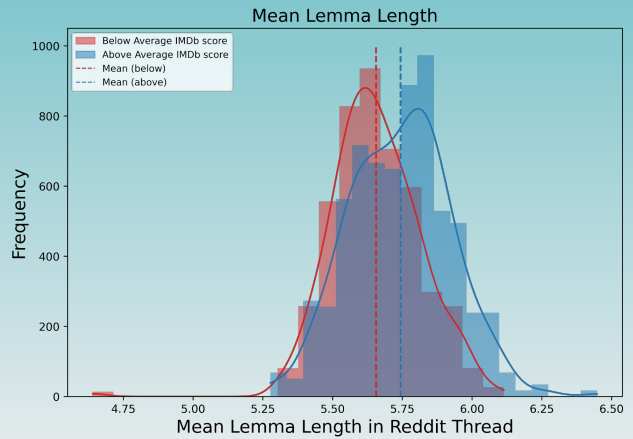
Data Understanding: NLP

Other features

- Proportion of positive comments
- Proportion of negative comments
- Average character count of comments
- Mean length of lemmas per discussion
- Proportion of unique lemmas per discussion
- Proportion of stop words

Examples:

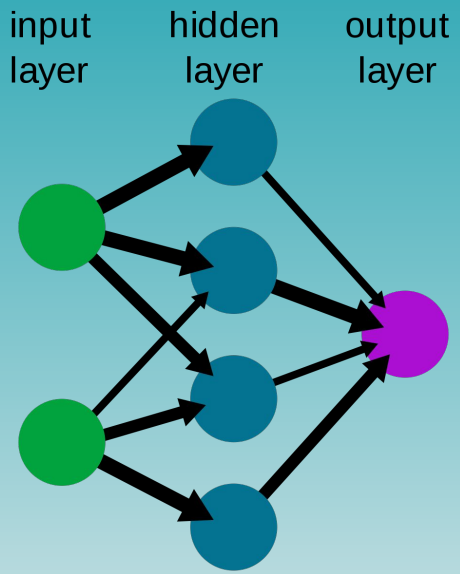




Modeling

Neural Networks:

- Machine learning practice mimicking the neurons of a brain. They can be tuned:
  - Inputs, layers, nodes, activation algorithms
  - Simple baseline model; 13 additional models
- Task: Regression
- Training, validation, and test datasets



Evaluation

Final Model

Test Performance

RMSE	On average, how far are predicted ratings from the true value?	0.57
R^2	How much of the variability in IMDb ratings does the model explain?	66.2%

Example Predictions

Title	Pred. IMdb Score	True IMDb Score
Onward	7.587624	7.4
Birds Of Prey	6.496331	6.0
Hacksaw Ridge	7.453154	8.1
The Irishman	7.711154	7.8
Big Bug	6.382244	5.5
The Legend Of Tarzan	6.075438	6.2
The Nun	5.296925	5.3
365 Days	4.796644	3.3
Hillbilly Elegy	6.538070	6.7
Lion	7.606026	8.0

Recommendations

- Use the model to predict IMDb scores
- Incorporate predicted score into decision-making
  - Bear in mind: Mean IMDb score is ~6.47
  - Prediction could be ~0.57 points off
- Split feedback by demographic (age, gender identity, etc.) to get more granular results
- If score is low, use qualitative feedback from comments
  - Scenes to reshoot, edit, or cut
  - Where to distribute the movie
  - How to market the movie

Limitations

- Only had discussions still available on Reddit (922 movies)
- Bias in movie selection
  - Major releases

Future work

- Experiment with different data
  - Letterboxd vs. Reddit
- Deployment: Python app & web app
- Apply model to YouTube comments on trailers

vs.

reddit

r/movies

Thank you

zshoorbajee@gmail.com

@zshoorbajee

linkedin.com/in/zshoorbajee/

zaid.fyi