DSTA Project
Ageflix - Movie Classification

Davit Melkonyan, Jakob Hennighausen, Leon Remke

Motivation

- Scenario

Goal

- Classify content ratings of movies
- Provide age-rating

Problems we hope to solve:
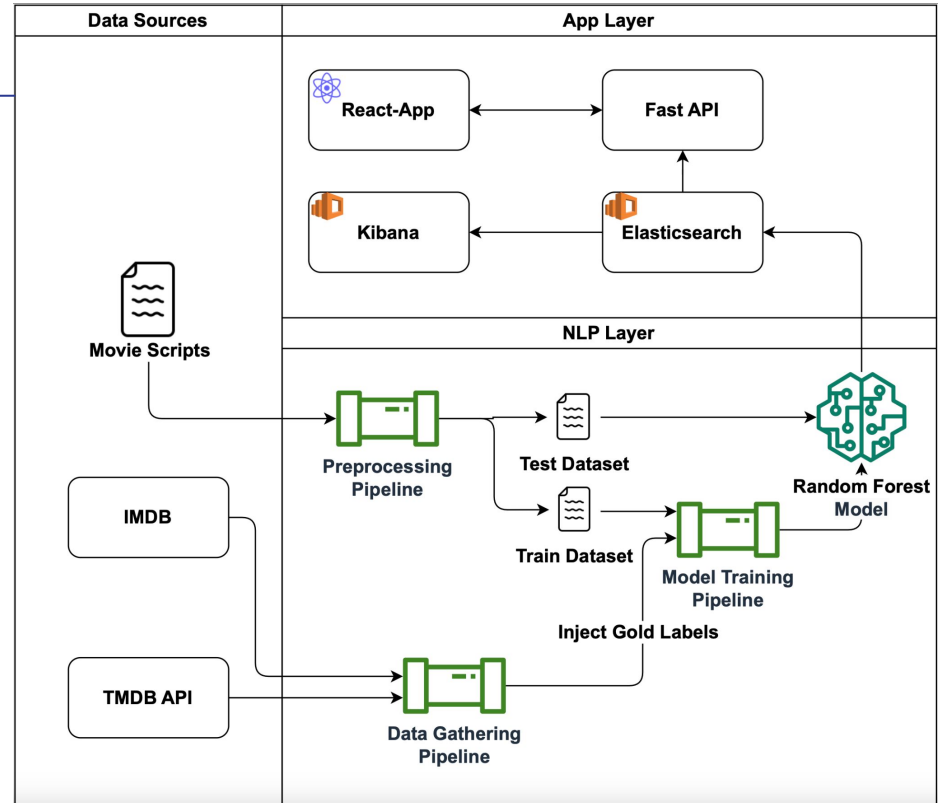
- Outdated Rating system

- Save time / effort

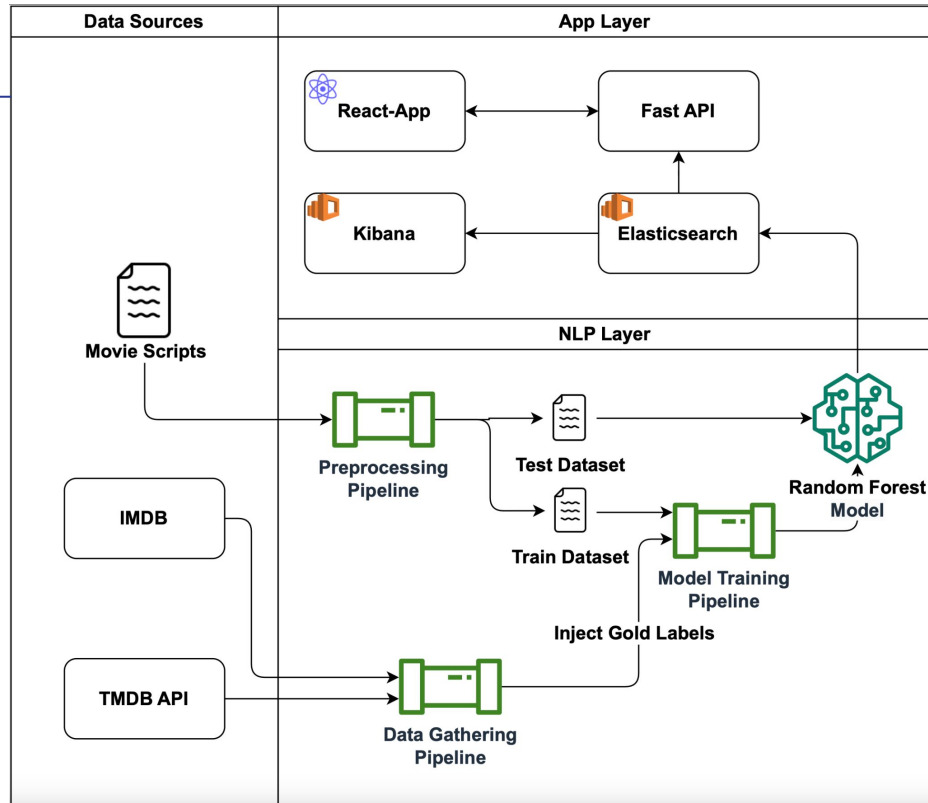- Spoiler-friendly

Approach

Data Gathering:

- Scripts from existing databank
- Age rating from TMDB API
- User votes for classification from Scraper libraries

[1] *"From None to Severe: Predicting Severity in Movie Scripts"* - Y. Zhang , M. Shafaei , F. González and T. Solorio
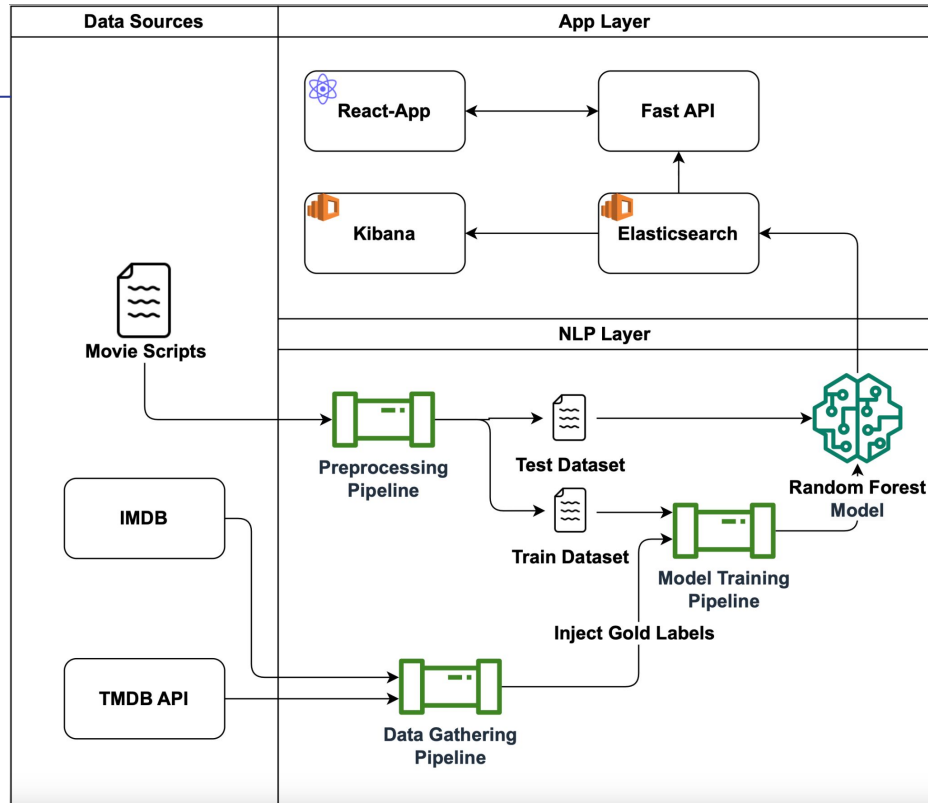
Approach

Data Preprocessing:

- clean up scripts (e.g. stop word removal, punctuation removal, etc)
- **Vectorization** of scripts as TF-IDF
- Merge into suitable **Dataframes**
- Split into **training** and **test** data

¹ *"From None to Severe: Predicting Severity in Movie Scripts"* - Y. Zhang , M. Shafaei , F. González and T. Solorio

Approach

Original Approach

- Mixture RNN-model from a paper[1]
- not feasible, due to large data size
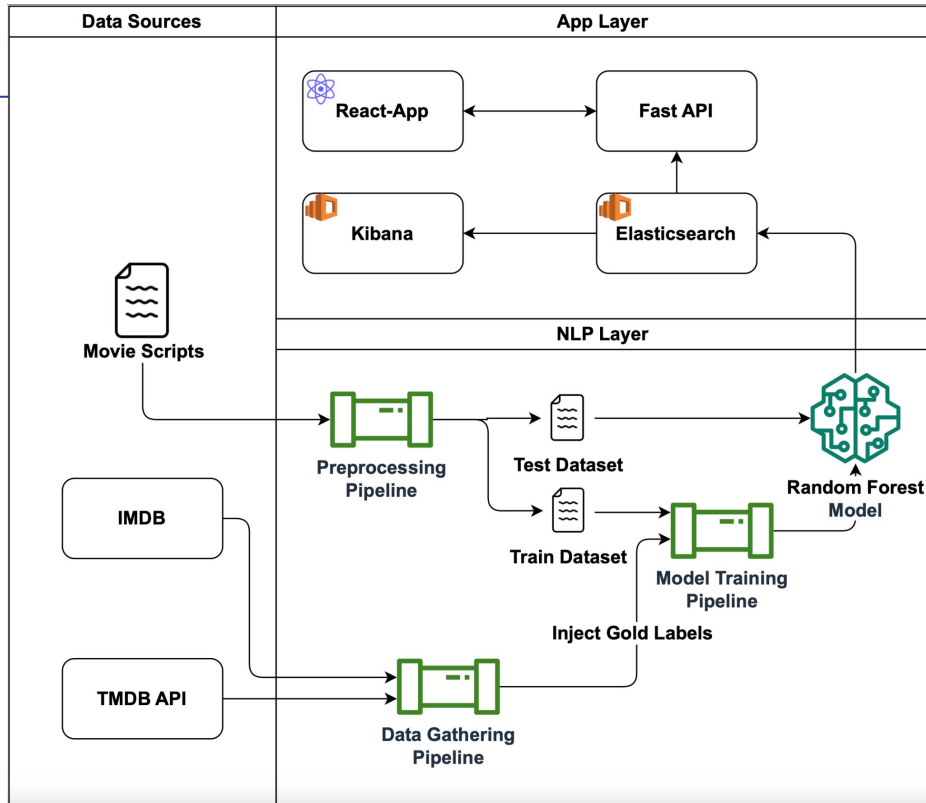- => very long runtime due to lack of processing power

[1] *"From None to Severe: Predicting Severity in Movie Scripts"* - Y. Zhang , M. Shafaei , F. González and T. Solorio

Approach

Our Alternative Approach

- Compare SVM to Random Forest model
- Build classifier
- Evaluation with a combination of basic heuristics (e.g. counts of profane words) and our model
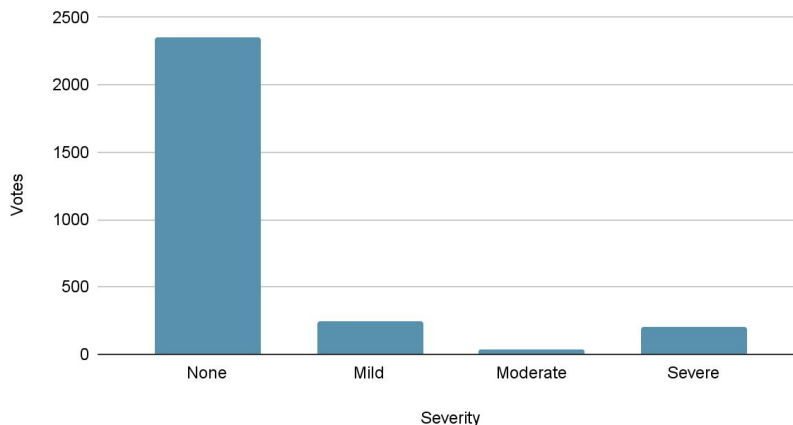
Data and Task

imdb_id	aspect	votes none	votes mild	votes moderate	votes severe	total votes	aspect rating	text
tt4154796	nudity	2351	251	33	207	2842	0	// script

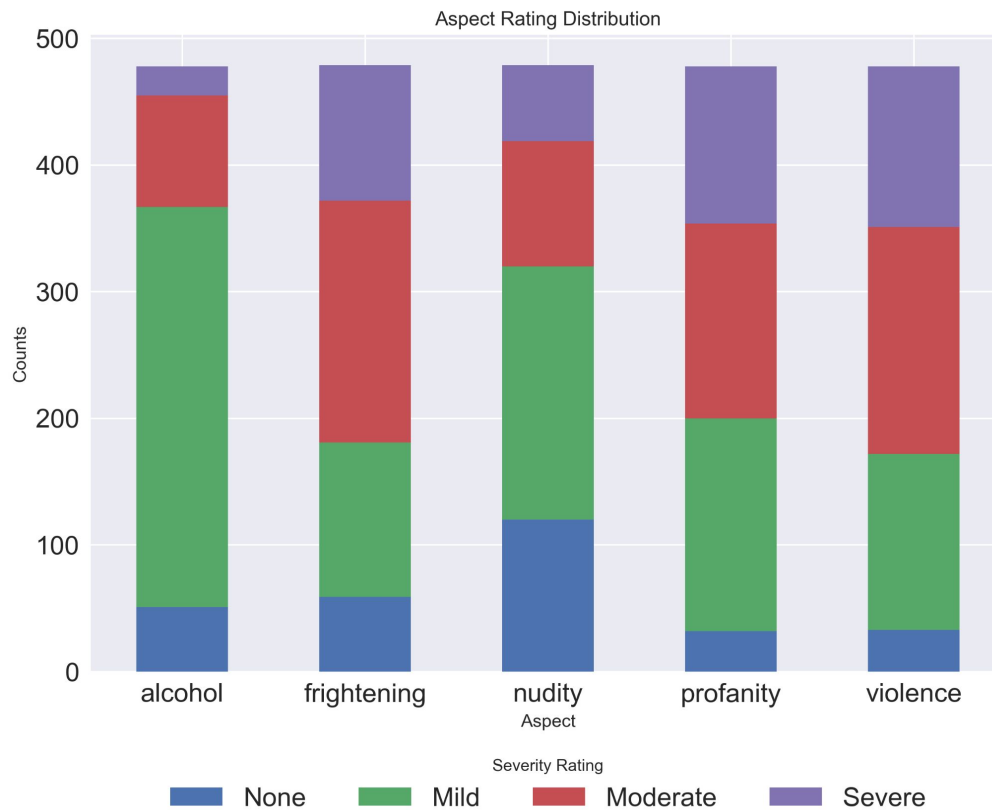tt4154796 = "Avengers Endgame"

Rating aspects, based on votes of users:

- Use of alcohol
- Frightening content
- Sex & nudity
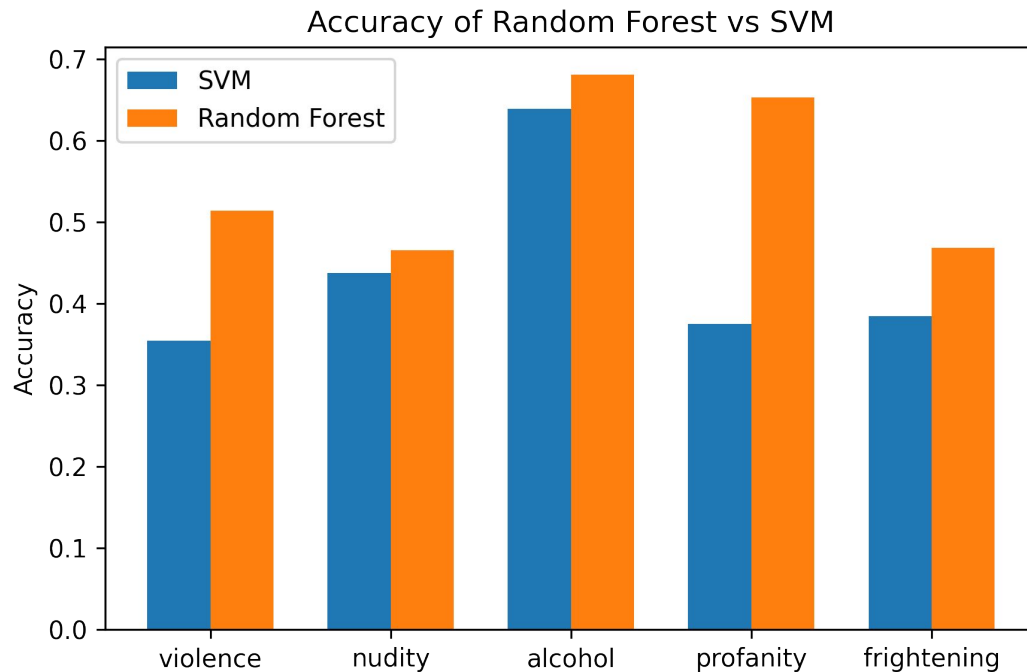- Use of profanity
- Violence

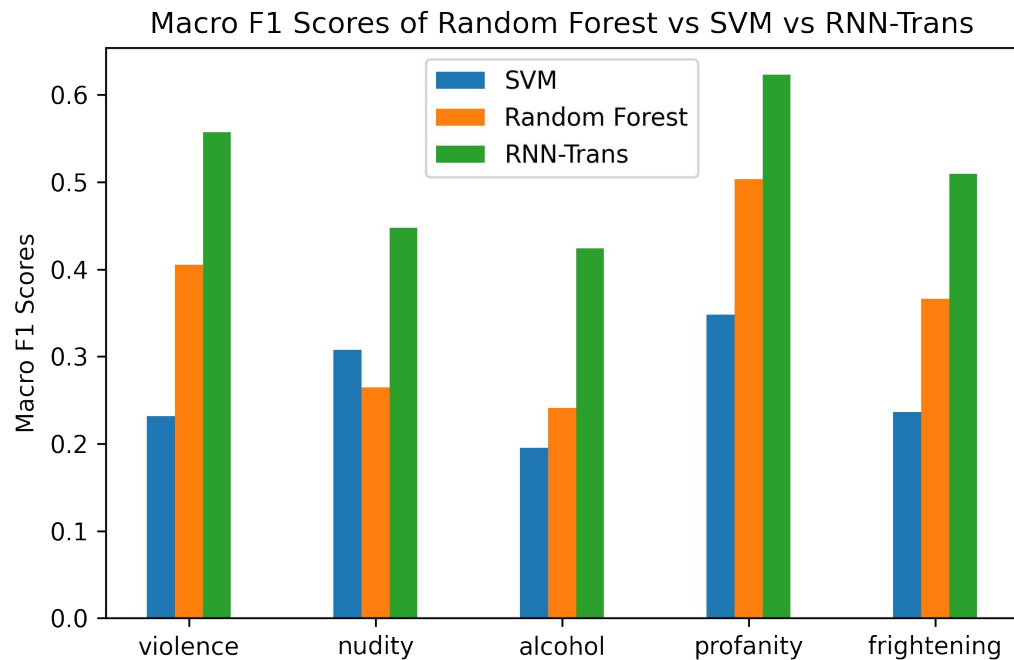Severity degree: nudity

Data and Task

Aspect Rating Distribution

Result

- Random Forest outperforms SVM
- Predict severity with accuracy up to 70%

Accuracy of Random Forest vs SVM

Result

Macro F1 Scores of Random Forest vs SVM vs RNN-Trans

¹ *"From None to Severe: Predicting Severity in Movie Scripts"* - Y. Zhang , M. Shafaei , F. González and T. Solorio

Conclusion & Outlook

- Our random forest model can predict severities almost accurately and give the appropriate age rating
- Advantages of our approach:
  - A lot faster than RNN approach
  - Based on user ratings from imdb -> Counteracts ratings creep[1]
- Possible improvements:
  - Add video and audio to input data
  - Differentiate between different types of movies (animated, non-animated, educational movie, etc.)
  - Differentiate between genre of movie (comedy)
  - Add more criteria to judge individual aspects (e.g. is the violence justified? Is the nudity sexual in nature?)[2]

[1] https://news.harvard.edu/gazette/story/2004/07/study-finds-ratings-creep/
[2] https://www.medialit.org/reading-room/whats-wrong-ratings