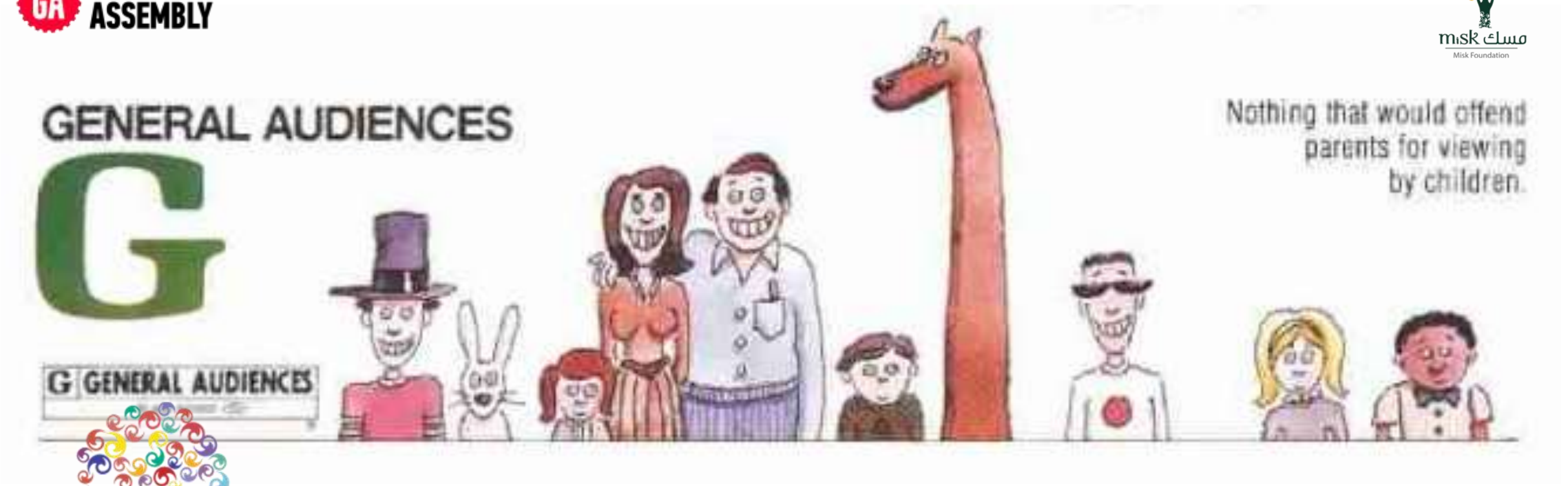


Predicting Content Rating of YouTube Movie Trailers based on Comments

By ohuod almarshdi

Outline

- Problem statement & objectives
- Data Sources
- EDA and Feature Engineering
- Data pre-processing
- Model
- Results
- Next steps

Problem statement & objectives

- Can we use YouTube trailer text feature to predict the category of movies ratings using classification models?

Why YouTube ?

- YouTube with unlimited upload capacity .
- You can find movies that have not yet been copyrighted and removed to watch .
- There’s billions of hours of content to watch.

Data sources

- Movie IMDb
- YouTube Movie Trailer
- YouTube Trailer Comments

Movie IMDb Data

For movie IMDb I have two datasets :

- IMDb API

1628 movie (not enough)

- IMDb Kaggle data

  5000 movie

YouTube trailer

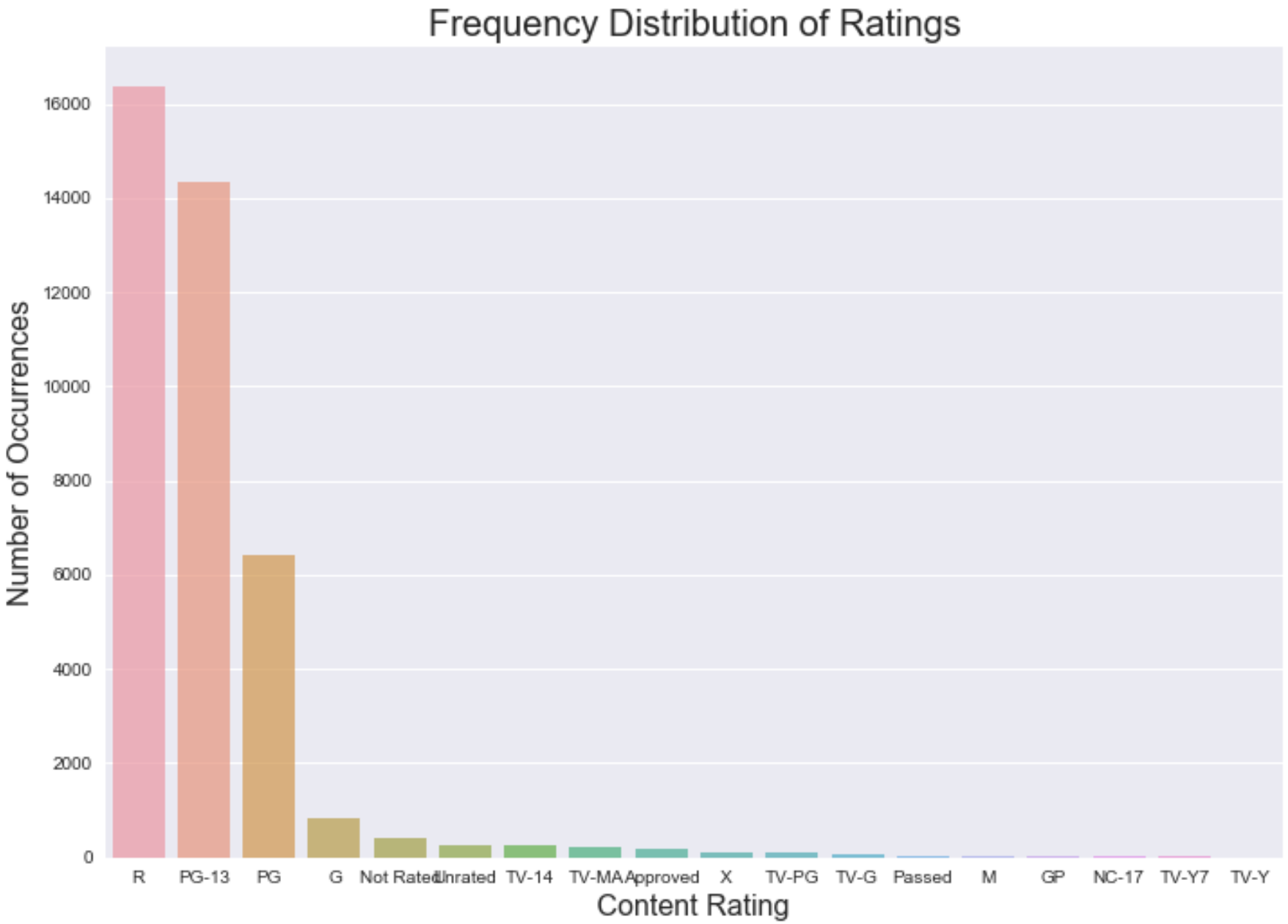
YouTube trailer Data

- YouTube API

  Movie YouTube dataset: 4147

  comments dataset : 34657

EDA


Frequency Distribution of Ratings
Number of Occurrences
16000
14000
12000
10000
8000
6000
4000
2000
0
R
PG-13
PG
G
Not Rated
Unrated
TV-14
TV-MA
Approved
X
TV-PG
TV-G
Passed
M
GP
NC-17
TV-Y7
TV-Y
Content Rating

Features Engineering

Transform Content Rating Classes into 3 categorize :

- Adults

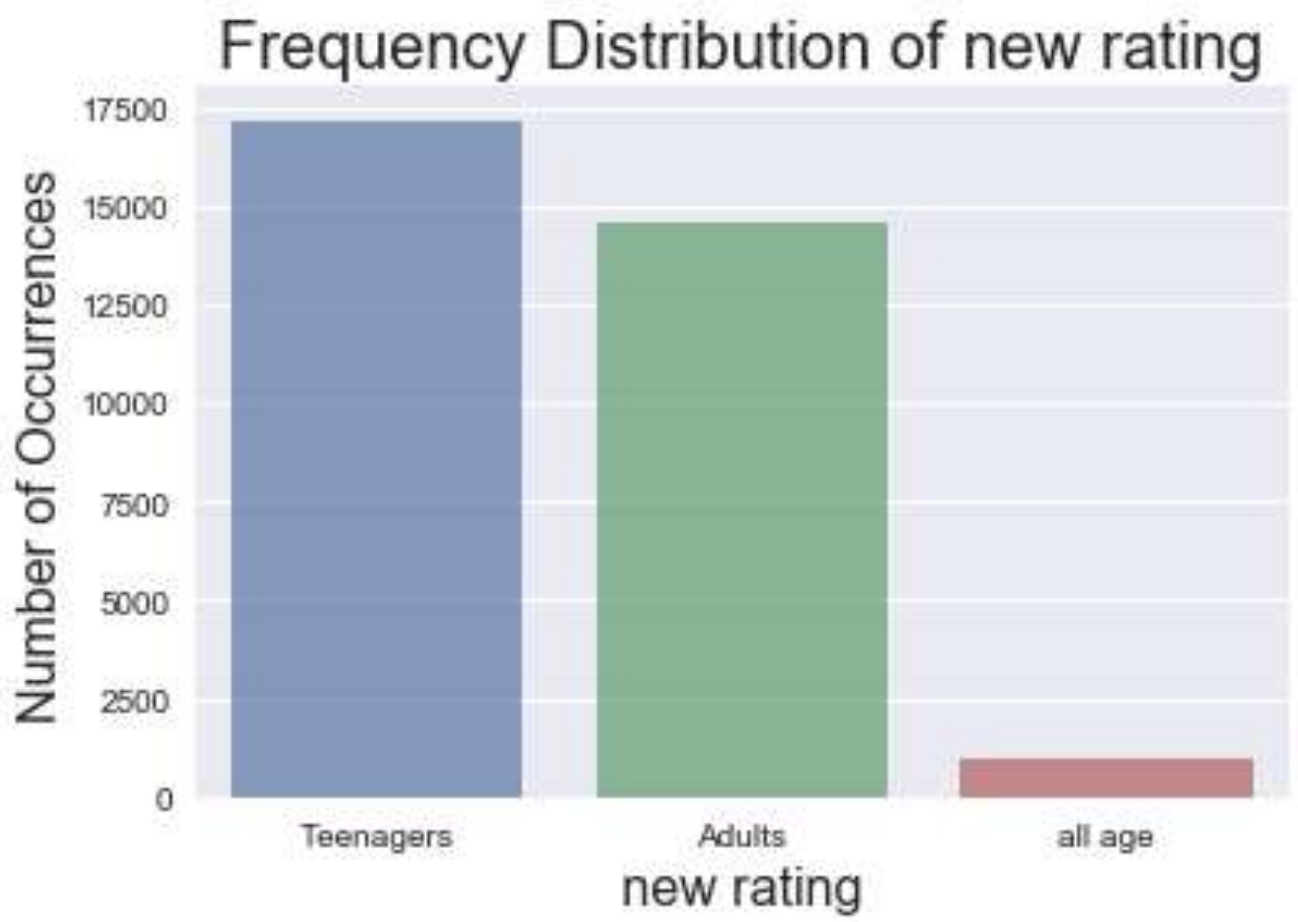
NC-17, TV-MA , X ,R

- Teenagers

PG, PG-13, TV-PG,TV-14

- All Ages

TV-Y7,TV-Y,TV-G,M,G ,GP



```
for txt in all_:
    sentences.append(txt.lower())
    tokenized = [t.lower().strip(":,.!?") for t in txt.split()]
    tokens.extend(tokenized)
    tokenizedSentences.append(tokenized)

hashtags = [w for w in tokens if w.startswith('#')]
ghashtags = [w for w in tokens if w.startswith('+')]
mentions = [w for w in tokens if w.startswith('@')]
links = [w for w in tokens if w.startswith('http') or w.startswith('www')]
```

Text cleaning

Remove

- non English words
- @ ,#,*,.?>/:;’\}[+-
- links and e-mails
- Emojis

Data pre-processing

- Natural language data pre-processing

Model Name	score
Linear SVC	0.529280
Logistic Regression	0.538835
Multinomial NB	0.540759
Random Forest Classifier	0.526347
LSTM	0.530

Model selection

Preliminary Results

- Multinomial NB

Multinomial NB

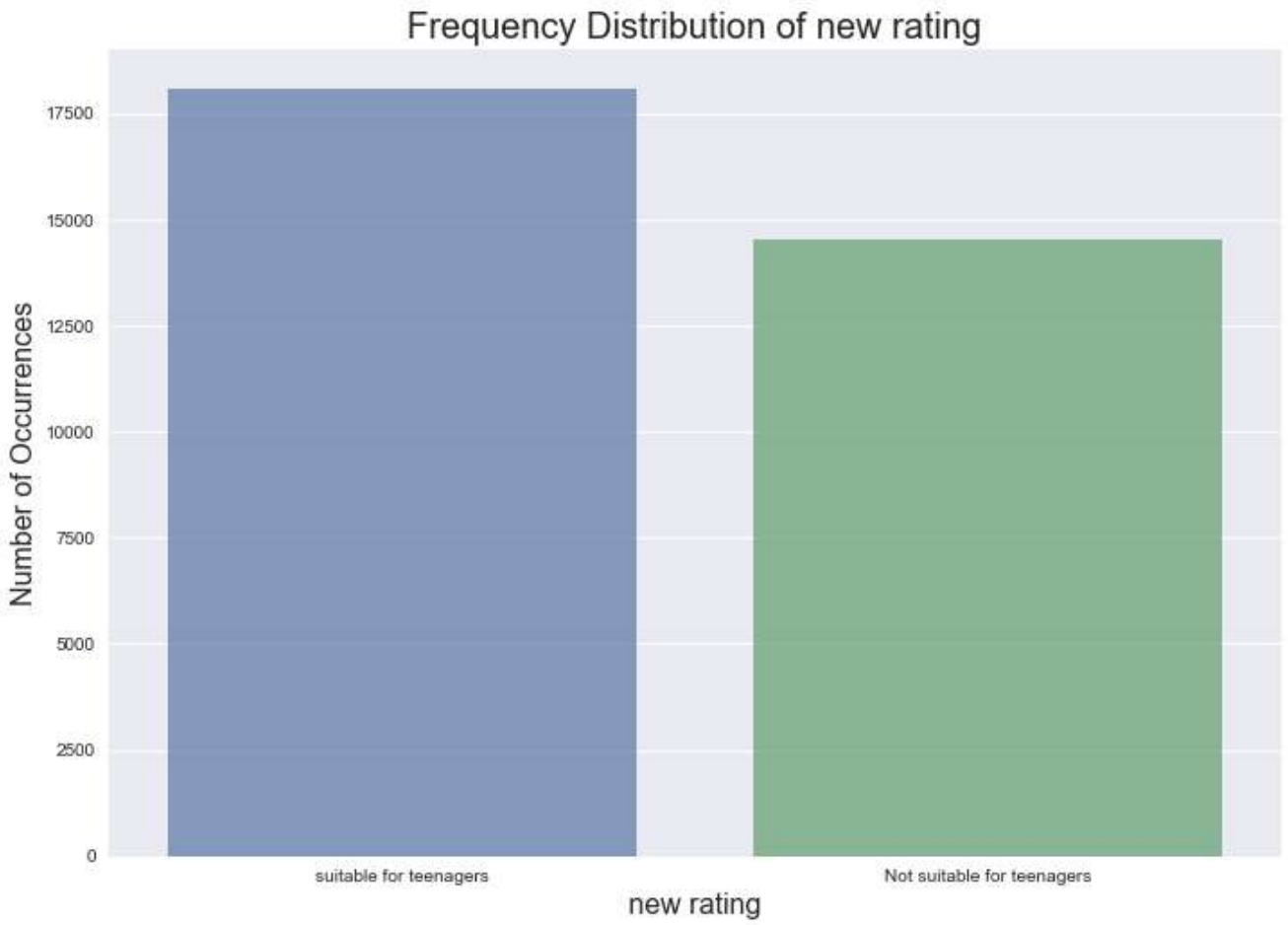
	precision	recall	f1-score	support
Teenagers	0.54	0.33	0.41	2813
all age	0.56	0.78	0.65	3337
Adults	0.00	0.00	0.00	193
avg / total	0.54	0.56	0.52	6343

- Multinomial Naive Bayes – Didn’t work well for 3 classes

What we can do to improve ?

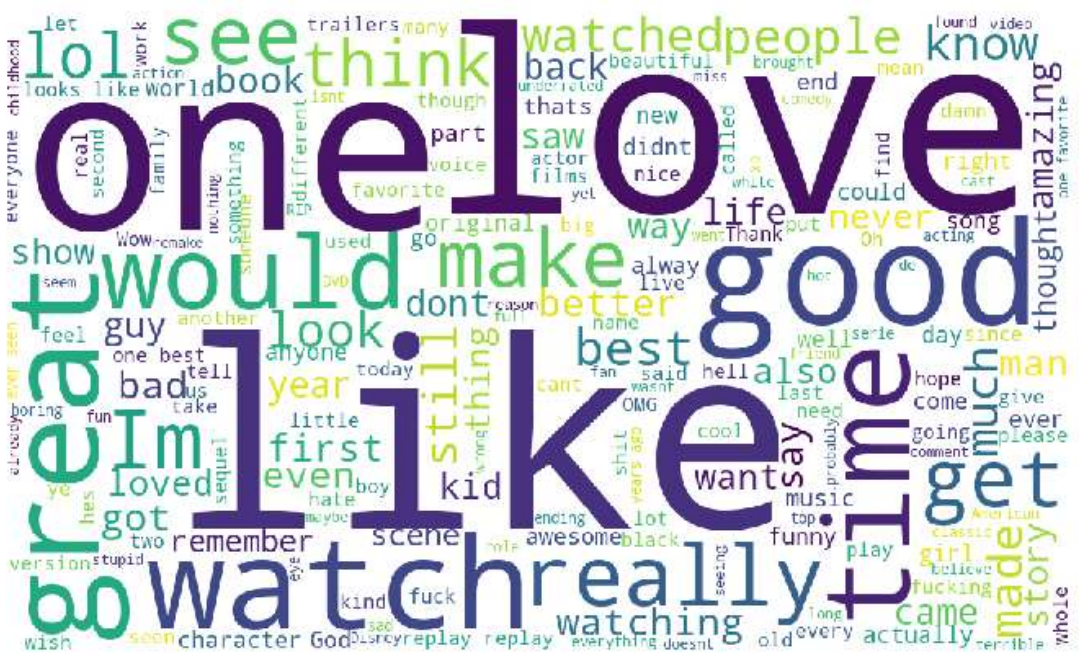
Can we minimize the class number for balance ?

- suitable for teenagers
- Not suitable for teenagers



Most common words in each

Not suitable for teenagers



suitable for teenagers

Model Name	score
Linear SVC	0.552269
Logistic Regression	0.563558
Multinomial NB	0.566523
Random Forest Classifier	0.555170
LSTM	0.567

Model selection

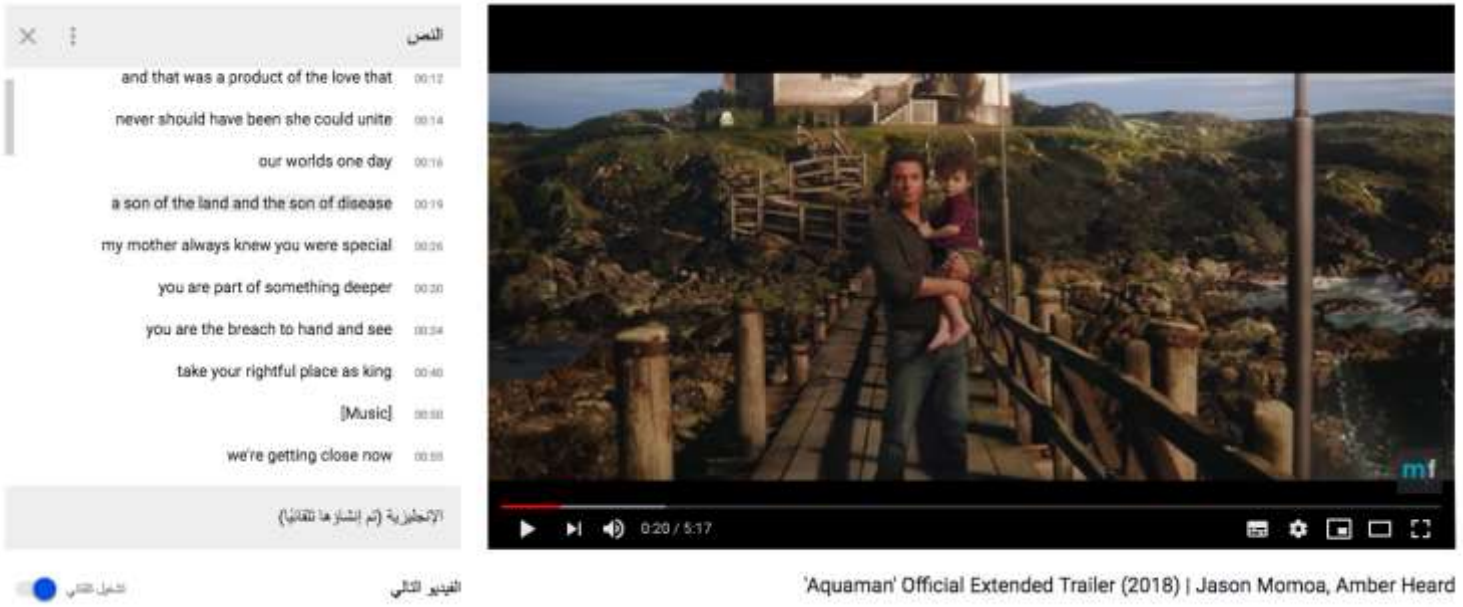
- LSTM
- Multinomial NB

Results

- Multinomial NB

	precision	recall	f1-score	support
Teenagers	0.56	0.29	0.38	2813
Adults	0.59	0.82	0.69	3530
avg / total	0.58	0.58	0.55	6343

- LSTM



Next steps

- Model with more comments and for more content ratings classes
- Work with video caption or transcript rather than comments