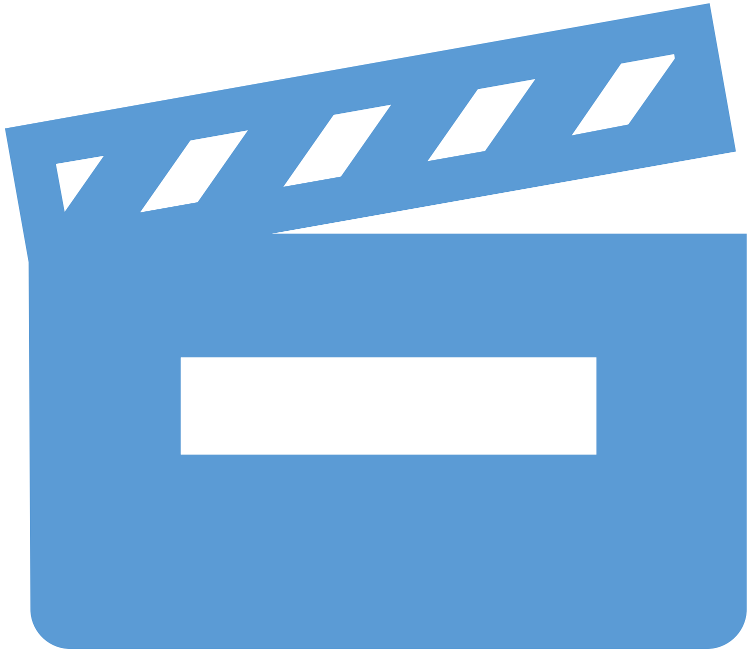
Aufteilung:

- Einführung: Problem, Motivation, KG, Daten-Übersicht Quelle Gabriel
- Daten-Analyse: Anna
- Statistische Modelle : Felix
- Transfomer: Matthias

FH University of Applied Sciences TECHNIKUM WIEN



Movie Plot Multi-label Genre Classification

Erlacher, Remta, Schachinger, Zechmeister

Problem Definition and Motivation

- **Problem:**
  - Classification of genres based on open source plot description text data.
  - Multilabel Classification
  - Irregular classification of film genres
- **Motivation:**
  - Best possible accuracy for test data
  - Possible application:
    - Better unification of classifications of film genres
    - Indie movie streaming service

Knowledge Gap and Research Question

- KG:
  - Comparison of the results of statistical models such as log-regression and transfer models. Which method gives better results on this data set?

- RQ:
  - How exactly can films be classified according to their genre using NLP techniques based on a short summary of the film's plot?

FH University of Applied Sciences TECHNIKUM WIEN

State-of-the-Art

Transformer-based Architectures:

- BERT
- GPT-3

Ensemble Models:

- Combine predictions from multiple models.

Transfer Learning and Fine-tuning:

- pre-trained models on large datasets and fine-tune them on specific tasks

Methodology Overview

Data
- Data Collection
- Plot Text Preprocessing
- Genre Clustering

Feature Extraction
- TF-IDF
- Tokenization

Model Training & Experiments
- Logistic Regression
- Pre-trained Model

Results & Evaluation
- Perfomance Measures
- Model comparison
- Conclusions

Data

- CMU Movie Summary corpus
  - Consists of 5 tsv or txt files with various movie information
  - Plot summaries obtained from Wikipedia
  - Metadata obtained from Freebase
  - Available at https://www.cs.cmu.edu/~ark/personas/
- Relevant information
  - 42.306 movie plot summaries obtained from Wikipedia
  - Metadata for 81.741 movies including the genres in json format

CMU Movie Summary Corpus

This page provides links to a dataset of movie plot summaries and associated metadata. This data was collected by David Bamman, Brendan O'Connor, and Noah Smith at the Language Technologies Institute and Machine Learning Department at Carnegie Mellon University.

Download

- Dataset [46 M] and readme: 42,306 movie plot summaries extracted from Wikipedia + aligned metadata extracted from Freebase, including:
  - Movie box office revenue, genre, release date, runtime, and language
  - Character names and aligned information about the actors who portray them, including gender and estimated age at the time of the movie's release
- Supplement: Stanford CoreNLP-processed summaries [628 M]. All of the plot summaries from above, run through the Stanford CoreNLP pipeline (tagging, parsing, NER and coref).

Further Reading

Please cite this paper if you write any papers involving the use of the data above:

- **Learning Latent Personas of Film Characters**
  David Bamman, Brendan O'Connor, and Noah A. Smith
  ACL 2013, Sofia, Bulgaria, August 2013

Acknowledgments

This research was supported in part by U.S. National Science Foundation grant IIS-0915187.

 For questions or comments, please contact David Bamman (dbamman@cs.cmu.edu).

Example

The following example illustrates the data and metadata available for *Indiana Jones and the Raiders of the Lost Ark*.

Movie metadata

Wikipedia movie ID	54166
Freebase movie ID	/m/0f4yh
Movie name	Indiana Jones and the Raiders of the Lost Ark
Movie release date	1981-06-12
Movie box office revenue	389925971
Movie runtime	115.0
Movie languages	Arabic Language, Nepali Language, Spanish Language, Hebrew Language, English Language, German Language
Movie countries	United States of America
Movie genres	Adventure, Costume Adventure, Action/Adventure, Action, New Hollywood, Airplanes and airports

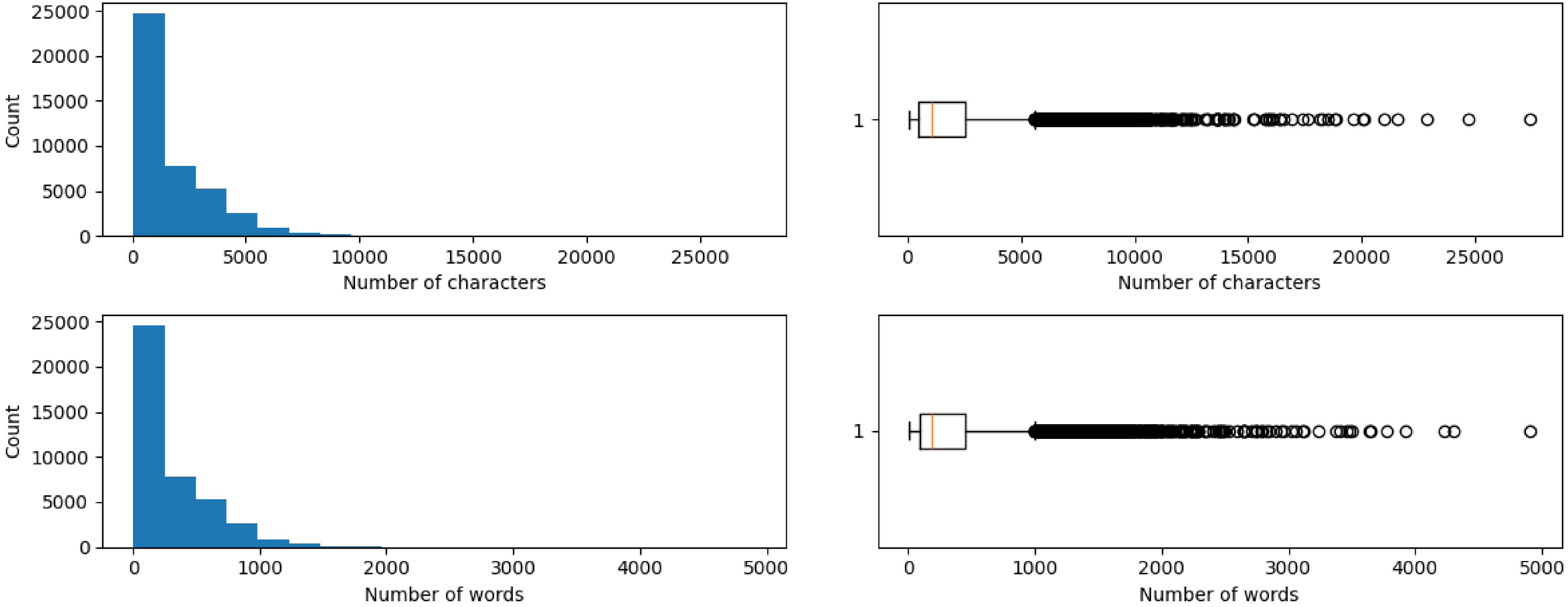
Character metadata

Wikipedia Movie ID	Freebase Movie ID	Character Name	Actor DOB	Actor gender	Actor height	Actor ethnicity	Actor Name	Actor age at movie release	Freebase character map
54166	/m/0f4yh	Dr. Marcus Brody	1922-05-31	M	1.816		Denholm Elliott	59	/m/02nwzzv
54166	/m/0f4yh	Simon Katanga	1949-10-20	M	1.87	/m/02w7gg	George Harris	31	/m/02nw_18
54166	/m/0f4yh	Dr. René Belloq	1943-01-18	M	1.77		Paul Freeman	38	/m/02nwzzg
54166	/m/0f4yh	Major Arnold Toht	1935-09-28	M			Ronald Lacey	45	/m/02nwzyz
54166	/m/0f4yh	Indiana Jones	1942-07-13	M	1.85	/m/01qhm_	Harrison Ford	38	/m/0k294p

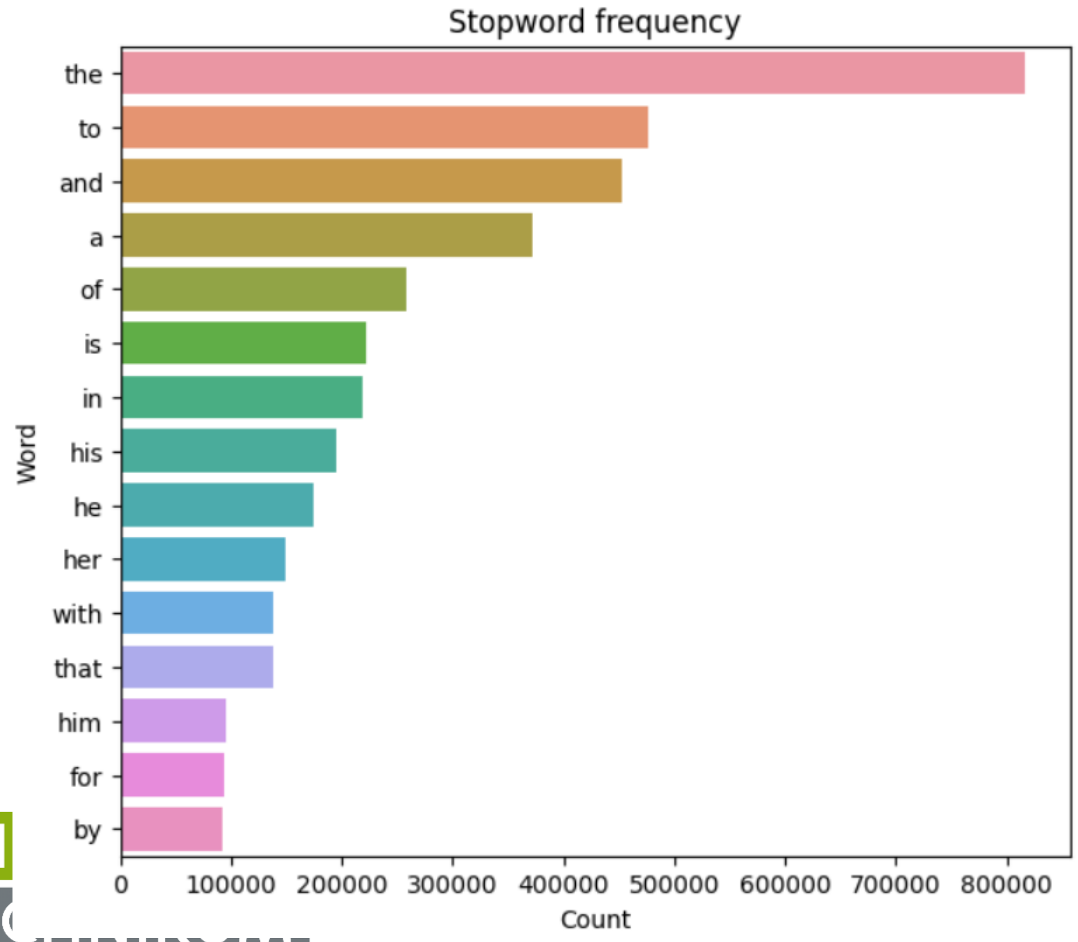
Data preprocessing

- Movie summaries
  - Converting text to lowercase
  - Removing Numbers, extra spaces and punctuations
- Genres
  - Converting genres to lists
  - Removing movies without assigned genres (411 movies lost)
- Merging movie and genre dataframes
  - 41793 merged movies

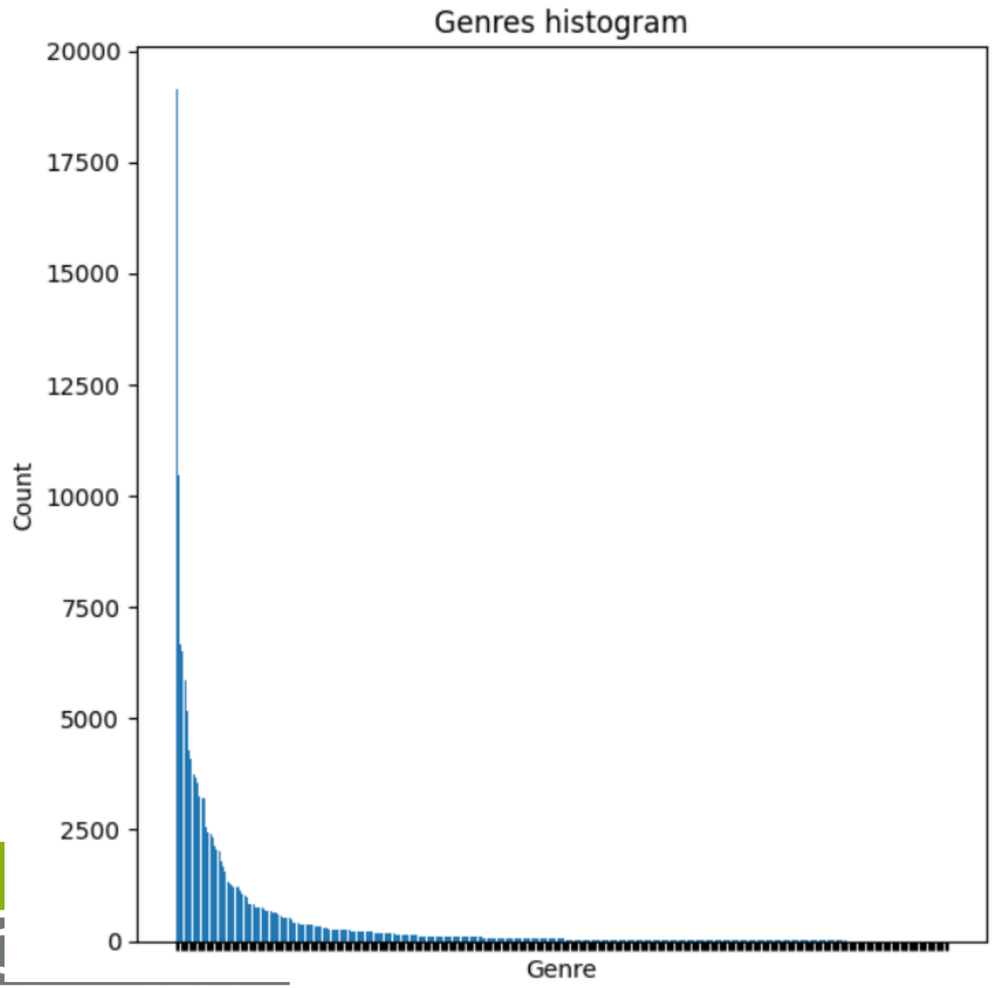
Data - Movie summaries

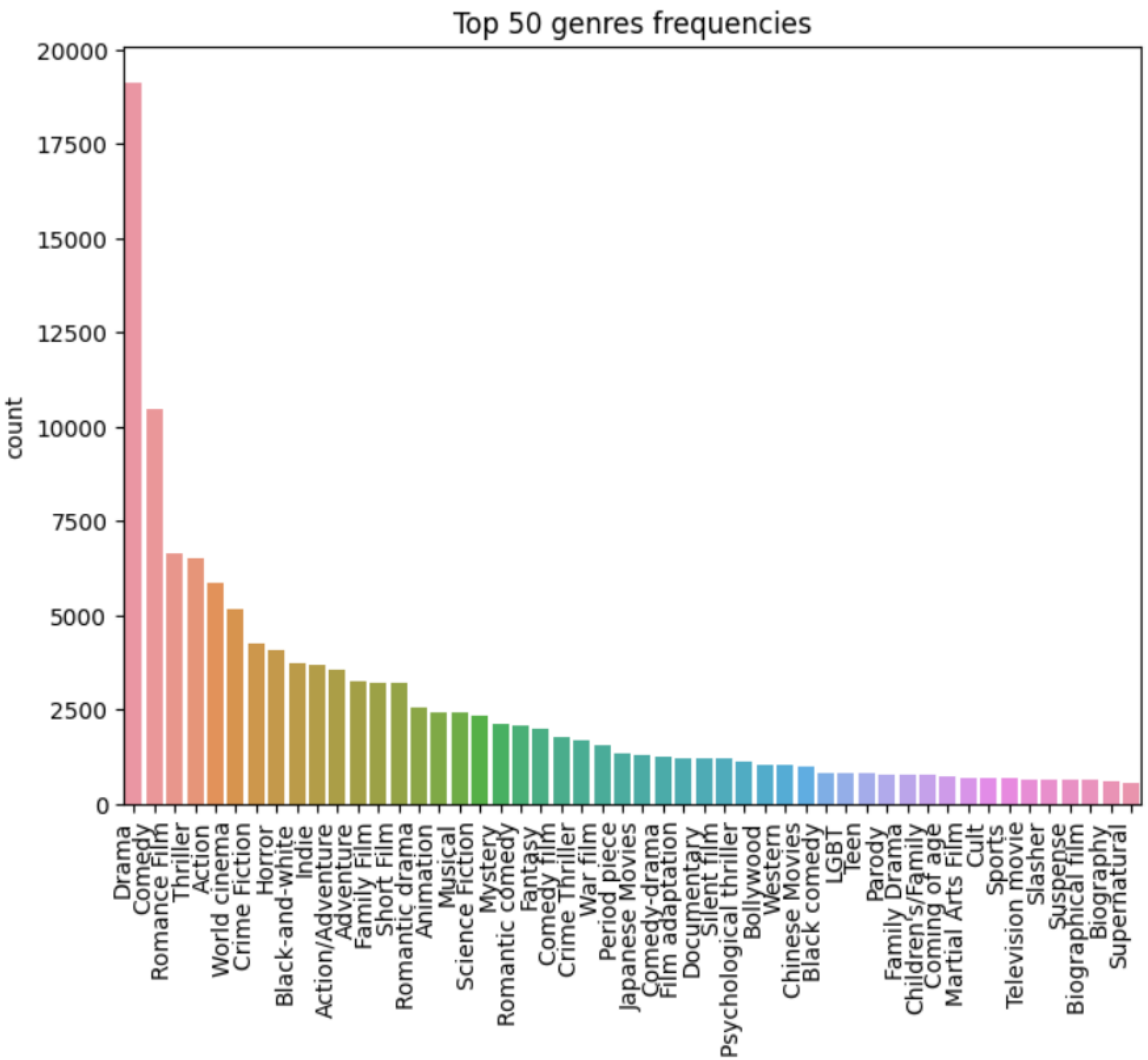


Data – Movie summaries



Data - Genres

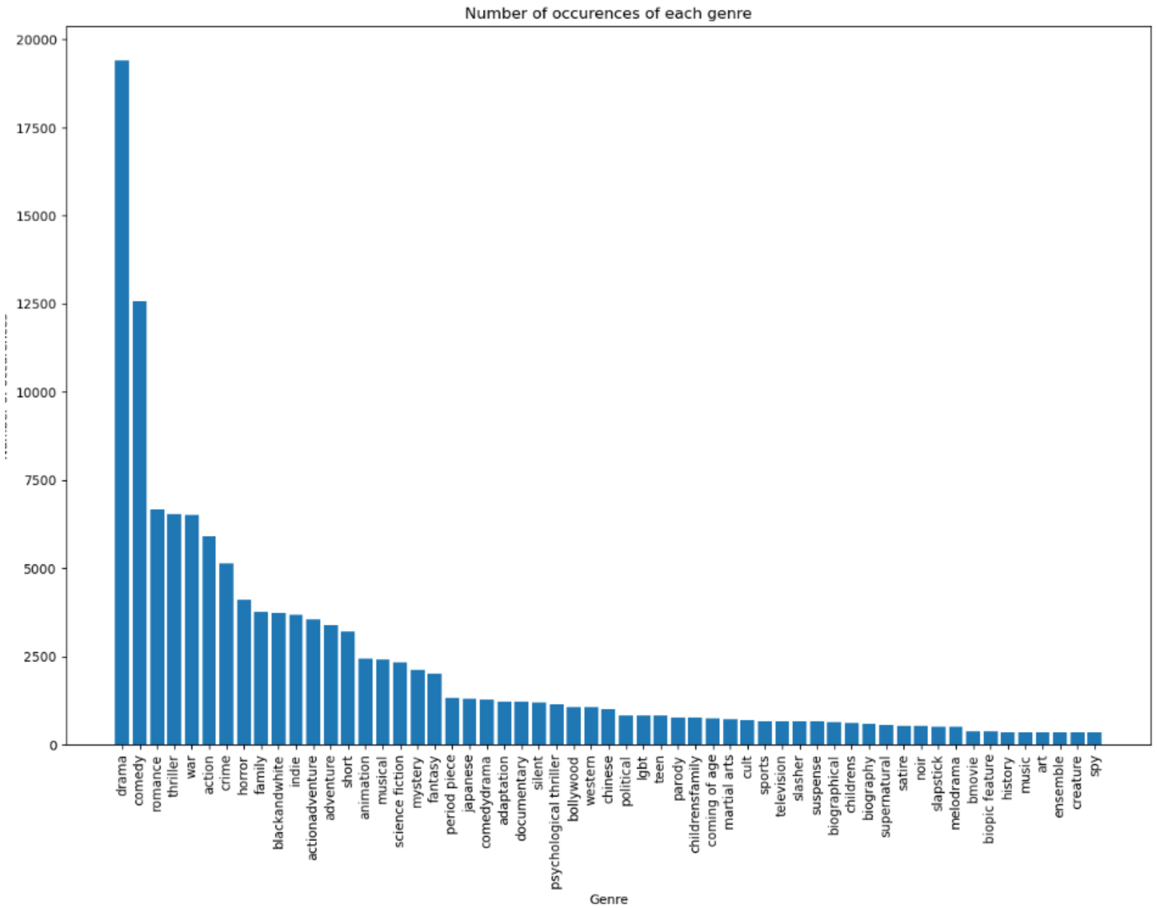

Genres histogram
Count
Genre
20000
17500
15000
12500
10000
7500
5000
2500
0

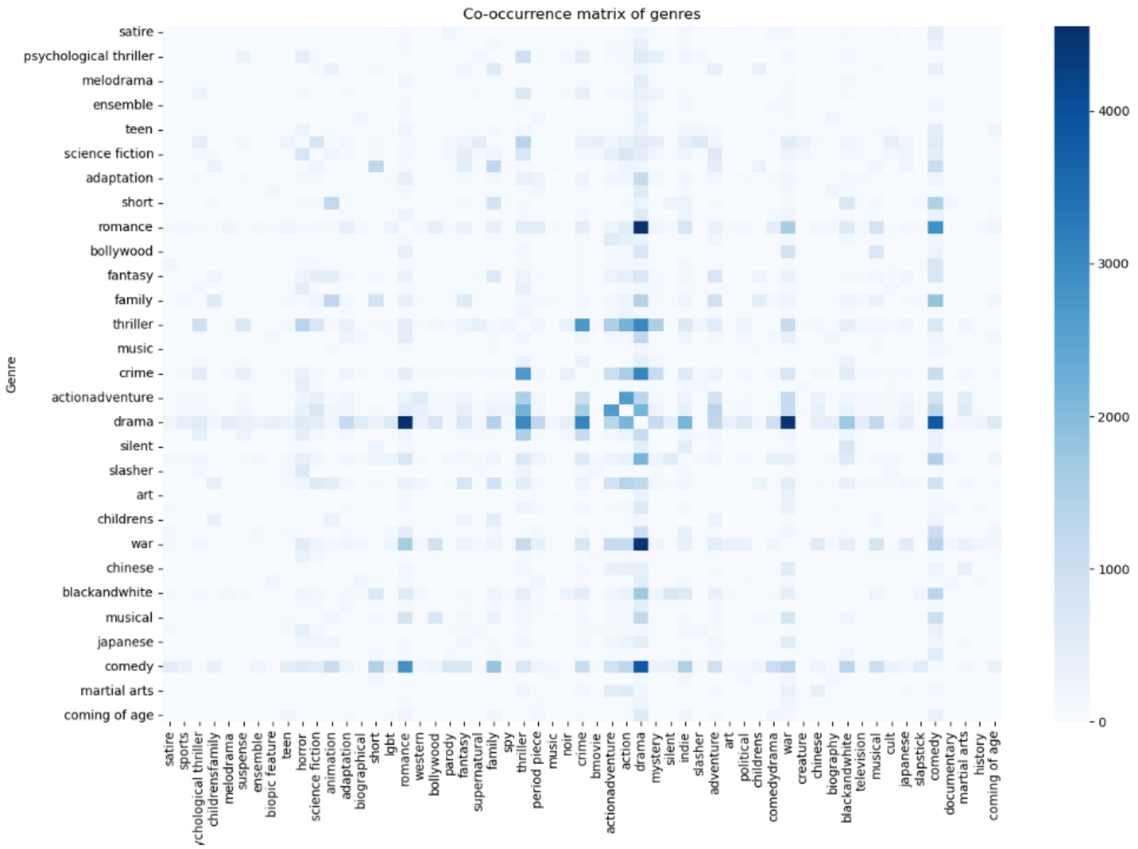

Top 50 genres frequencies
count
20000
17500
15000
12500
10000
7500
5000
2500
0
Drama
Comedy
Romance Film
Thriller
Action
World cinema
Crime Fiction
Horror
Black-and-white
Indie
Action/Adventure
Adventure
Family Film
Short Film
Romantic drama
Animation
Musical
Science Fiction
Mystery
Romantic comedy
Fantasy
Comedy film
Crime Thriller
War film
Period piece
Japanese Movies
Comedy-drama
Film adaptation
Documentary
Silent film
Psychological thriller
Bollywood
Western
Chinese Movies
Black comedy
LGBT
Teen
Parody
Family Drama
Children's/Family
Coming of age
Martial Arts Film
Cult
Sports
Television movie
Slasher
Suspense
Biographical film
Biography
Supernatural

Additional Data Cleaning – Removing Movies with minor genres

- 363 genres in original data
- TDF-IDF + KNN 15 categories -> 14 usefull 1 rest
- Reduced to 57 genres
- Deleted all with less than 341 occurences
- 41793 to 41549 entries

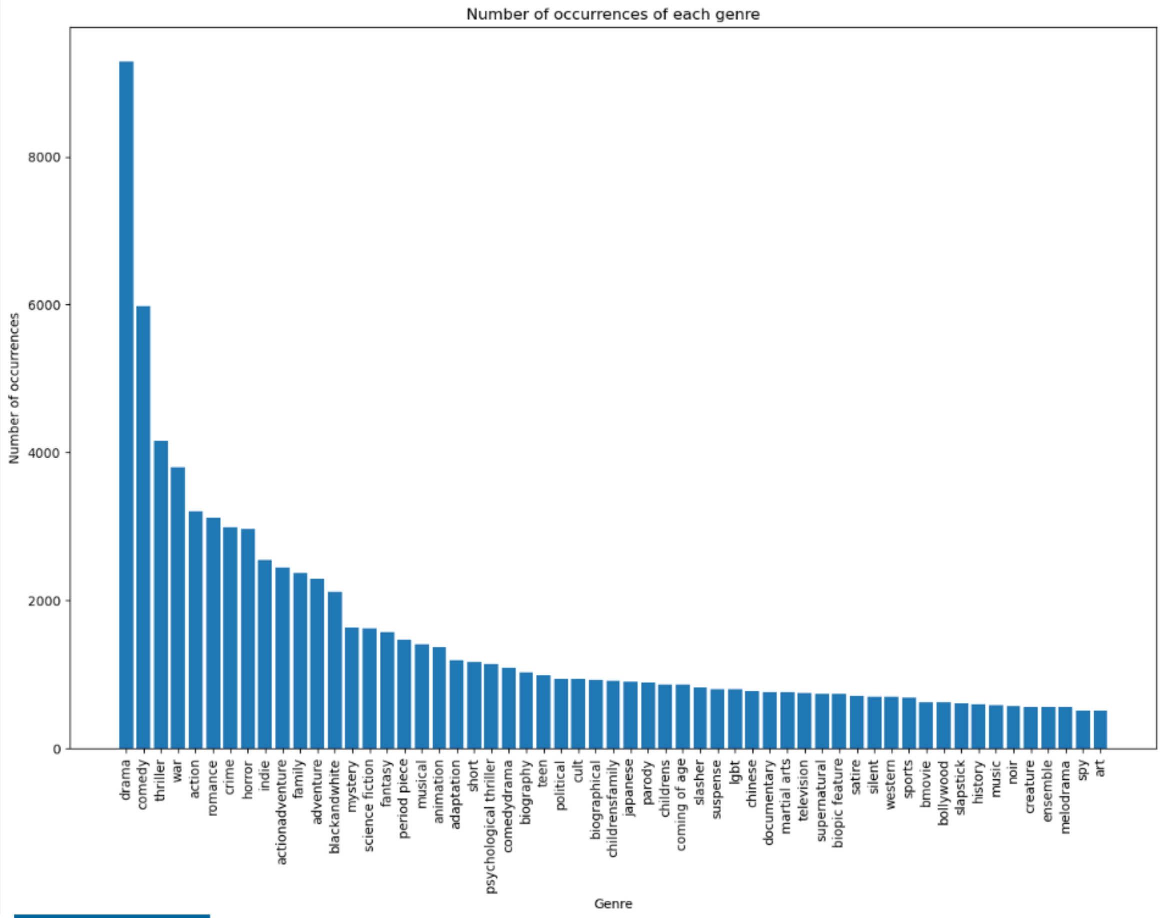
Additional Data Cleaning

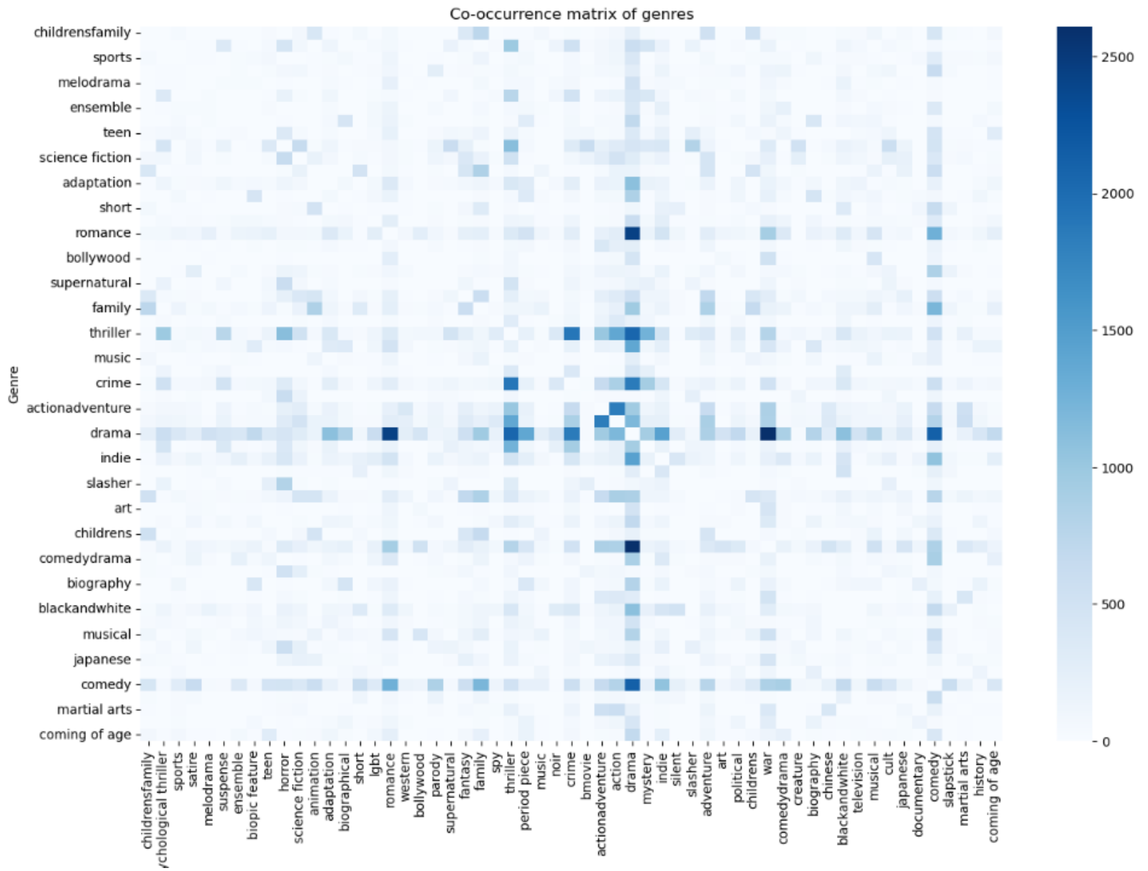




Additional Data Cleaning & balancing

Take random sample with number of least frequent genres from dataset for each genre





WIEN

Output Model Example: Log-Reg-Reduced

```
for i in range(5):
    k = xval.sample(1).index[0]
    print("Movie: ", df_test['title'][k], "\nPredicted genre: ", new_val(xval[k]))
    print("Actual genre: ",df_test['genre'][k], "\n")
```

[69] ✓ 0.0s

```
Movie:  Jill Rips
Predicted genre:  [('crime', 'drama', 'thriller')]
Actual genre:  ['thriller', 'crime']

Movie:  Forces of Nature
Predicted genre:  [('comedy',)]
Actual genre:  ['comedy', 'romance']

Movie:  Lifeforce
Predicted genre:  [('horror', 'science fiction')]
Actual genre:  ['science fiction', 'horror', 'adventure', 'cult']

Movie:  The Golden Blade
Predicted genre:  [('action', 'adventure', 'fantasy')]
Actual genre:  ['action', 'adventure']

Movie:  Heist
Predicted genre:  [('crime',)]
Actual genre:  ['thriller', 'crime', 'drama']
```

Multi-Label Logistic Regression Model - Performance

Before additional data cleaning

- 57 categories
- ~42000 entries
- Highest/lowest category ~ 56

	precision	recall	f1-score	support
micro avg	0,716	0,267	0,389	25598
macro avg	0,466	0,102	0,149	25598
weighted avg	0,637	0,267	0,341	25598
samples avg	0,552	0,311	0,361	25598

After additional data cleaning

- 57 categories
- ~19500 entries
- Highest/lowest category ~18

	precision	recall	f1-score	support
micro avg	0,799	0,261	0,393	17084
macro avg	0,745	0,142	0,213	17084
weighted avg	0,775	0,261	0,342	17084
samples avg	0,646	0,262	0,347	17084

FH University of Applied Sciences TECHNIKUM WIEN

Naive Bayes

Reduced:

	precision	recall	f1-score	support
micro avg	0,335	0,694	0,452	17084
macro avg	0,335	0,639	0,430	17084
weighted avg	0,367	0,694	0,468	17084
samples avg	0,414	0,655	0,459	17084

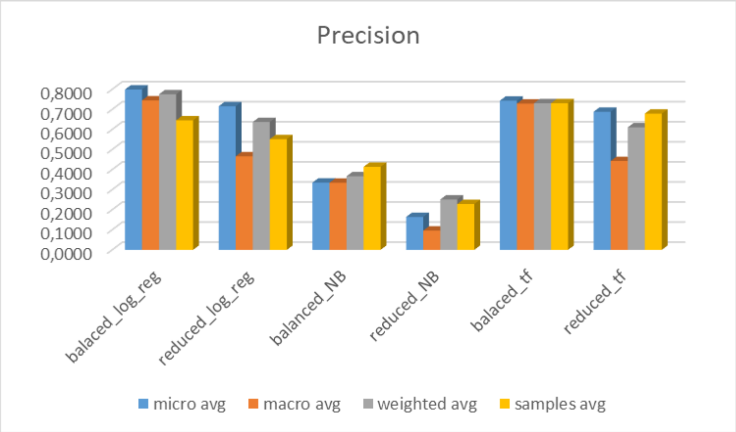
Balanced:

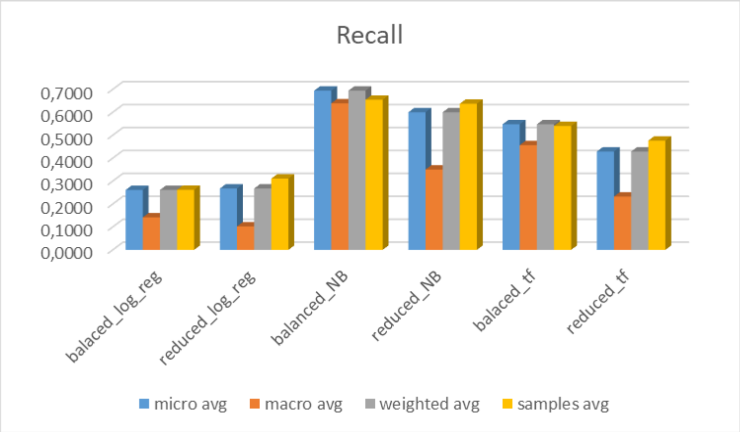
	precision	recall	f1-score	support
micro avg	0,164	0,600	0,258	25598
macro avg	0,096	0,350	0,144	25598
weighted avg	0,251	0,600	0,333	25598
samples avg	0,228	0,637	0,290	25598

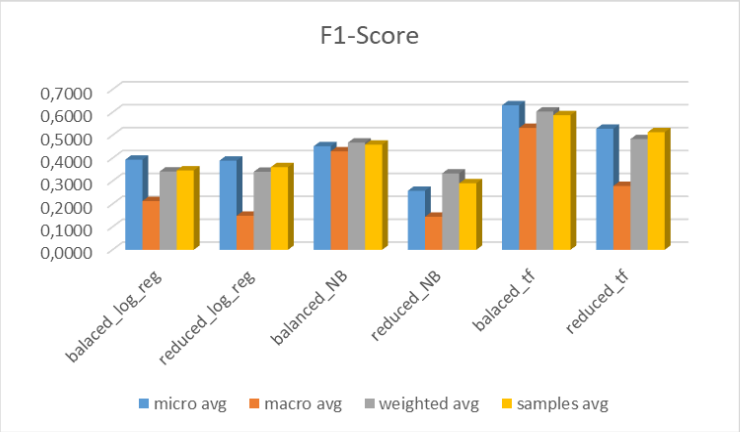
SVM

Does not converge

Result comparison







Transformer

- Base model: DistilBERT[1]
- Tokenizer: DistilBertTokenizerFast (max 512 token)
- Fine-tuned on reduced dataset for 6 epochs[2]
- Fine-tuned on balanced dataset for 8 epochs[3]

1. distilbert-base-uncased 2. matthiasr/genre_pred_model_reduced 3. matthiasr/genre_pred_model_balanced

Transformer Results

Reduced	Precision	Recall	F1-score
Micro average	0.69	0.43	0.53
Macro average	0.44	0.23	0.28
Weighted average	0.61	0.43	0.48
Sample average	0.68	0.48	0.51

Balanced	Precision	Recall	F1-score
Micro average	0.74	0.55	0.63
Macro average	0.73	0.46	0.53
Weighted average	0.73	0.55	0.60
Sample average	0.73	0.54	0.59

FH University of Applied Sciences
TECHNIKUM
WIEN