Predicting Movie Genres based on IMDB Descriptions

Sebastian Pinter, Daniel Chudý, Filip Faber, Tim Dirr

np.nan

194.093 Natural Language Processing and Information Extraction 2024/25

TU Wien, Austria

Agenda

- The Task
- Data
- Baseline
- Identified Issues
- Improvements & Results
- Further Work

Genre Prediction

Predict movie genres based on textual description

- Text Classification
- Multilabel
- Hard Task!

Based on this:

"A Pink/Roman porno with a yakuza character or two"

Predict this:

["Action", "Crime"]

Data

IMDB movie dataset containing genres (ground truth), textual movie plot descriptions and imdb-id

- ~190k rows, varying description lengths
- some rows with no description
- lots of rows where description is cut of
  (“This movie talks about the ….”)
- some nonsense plot descriptions
  (“Add a plot”), (“plot unknown”), (“under wraps”)
- heavy class imbalance

Basic preprocessing

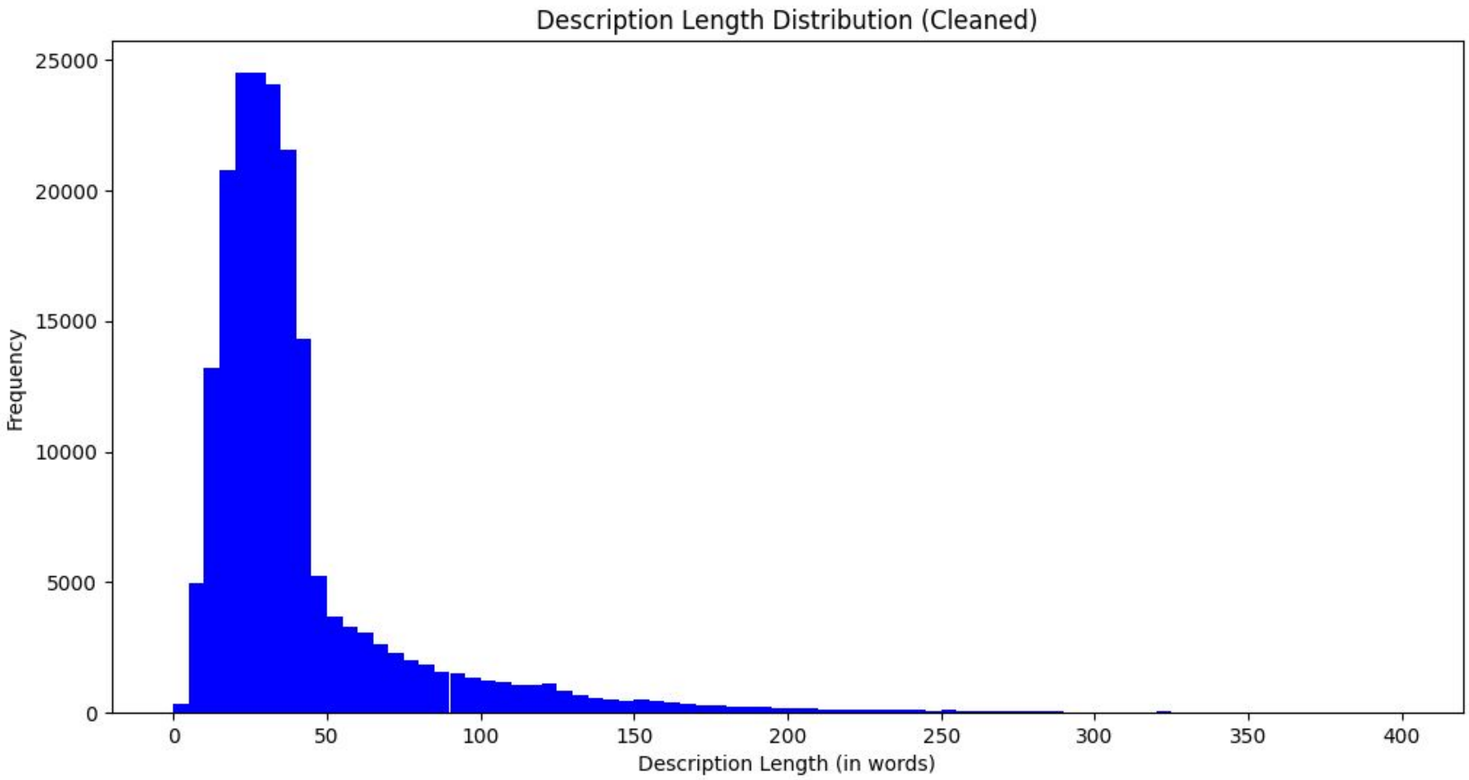
We removed nonsense description by pattern matching

Still, lots of missing and incomplete description

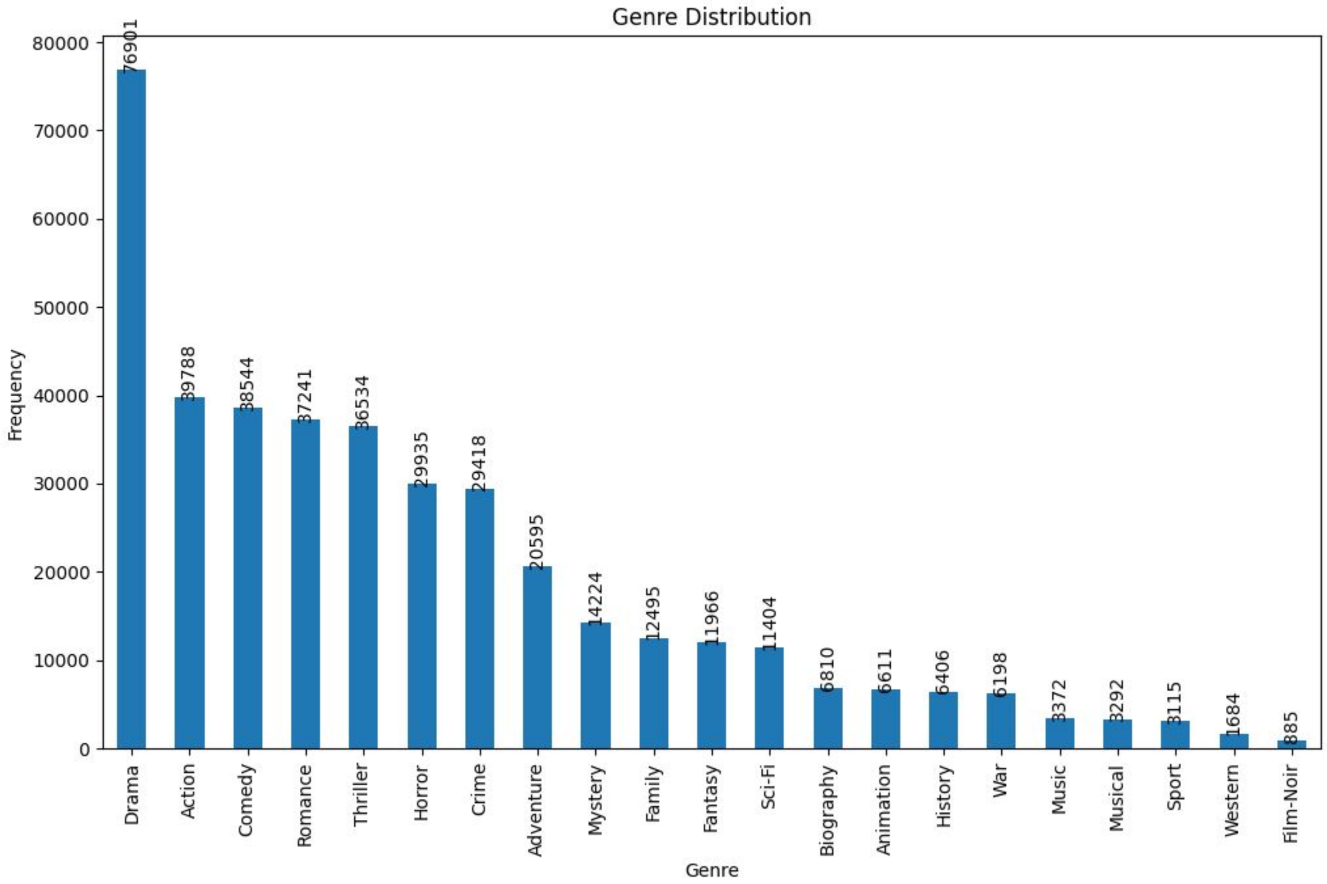
$\longrightarrow$ crawl IMDB database

Lemmatization

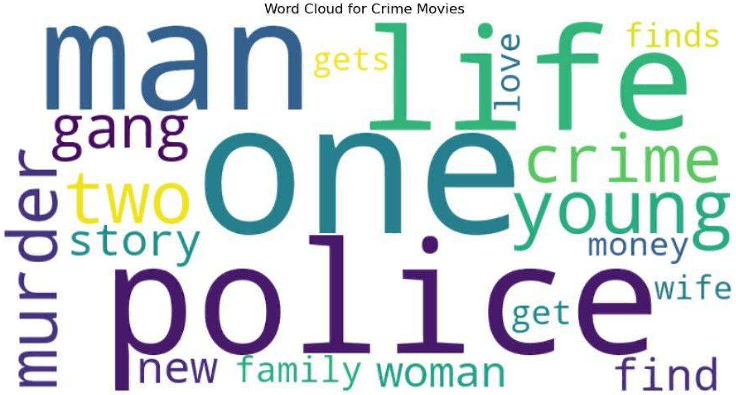
Data - Description Length

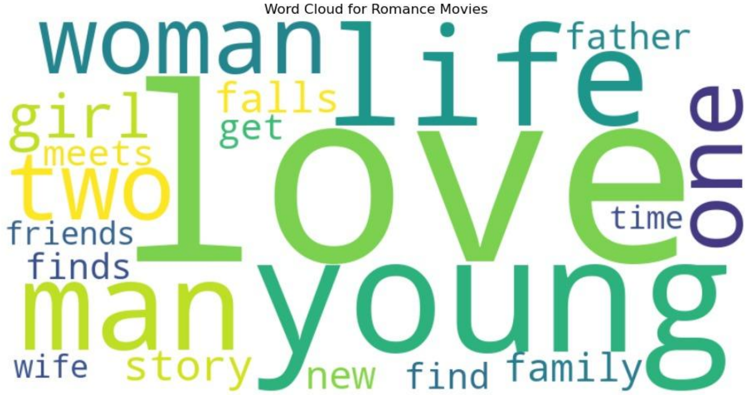


Data - Genre Distribution



Data - Word Clouds





BREAK: SWITCH FROM A TO B

Baseline - Models

Text Modelling:

Bag Of Words: Count / Tf-Idf

Classifier:

- Multilabel Classifier -> Training one clf per class
- Logistic Regression, KNN, Decision Tree, ...

Keeping it simple!

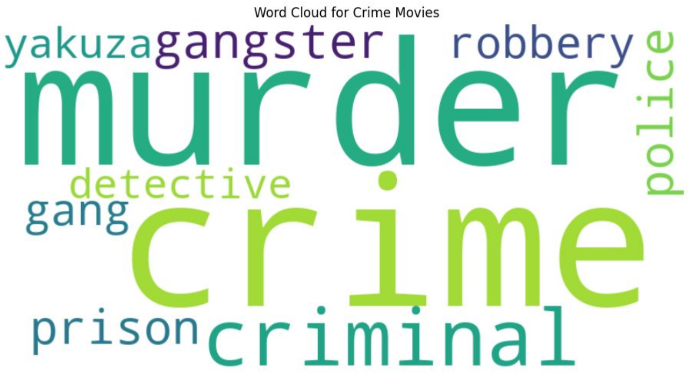
Baseline - Metrics

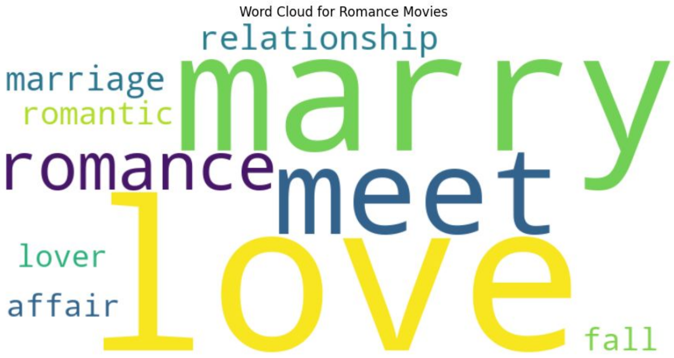
CLF	BoW	Jaccard	Hamming	Prec.	Recall	at-least-1	at-least-2
LReg	**Count**	0.29	0.09	0.47	0.35	0.57	0.14
LReg	**Tf-Idf**	0.33	0.09	0.52	0.38	0.62	0.16

$$\text{Jaccard Score}\ (\hat{y}, y) = \sum_{i=0}^{n_{samples}-1} \frac{1}{n_i} \frac{|\hat{y}_i \cap y_i|}{|\hat{y}_i \cup y_i|} \quad \text{Intersection over Union per Sample}$$

$$\text{Hamming Loss}\ (\hat{y}, y) = \frac{1}{n_{samples} \cdot n_{labels}} \sum_{i=0}^{n_{samples}-1} \sum_{j=0}^{n_{labels}-1} \mathbf{1}(\hat{y}_{i,j} \neq y_{i,j}) \quad \text{Fraction of wrong predicitons}$$

Feature Importance





BREAK: SWITCH FROM B TO DANIEL

Deep Learning Model

- DistilBERT (40% faster)
- dataset of 17k rows (0.8/0.1/0.1 split).
- 3 epochs
- probability threshold (0.4 … 0.5)

DL vs. LogReg

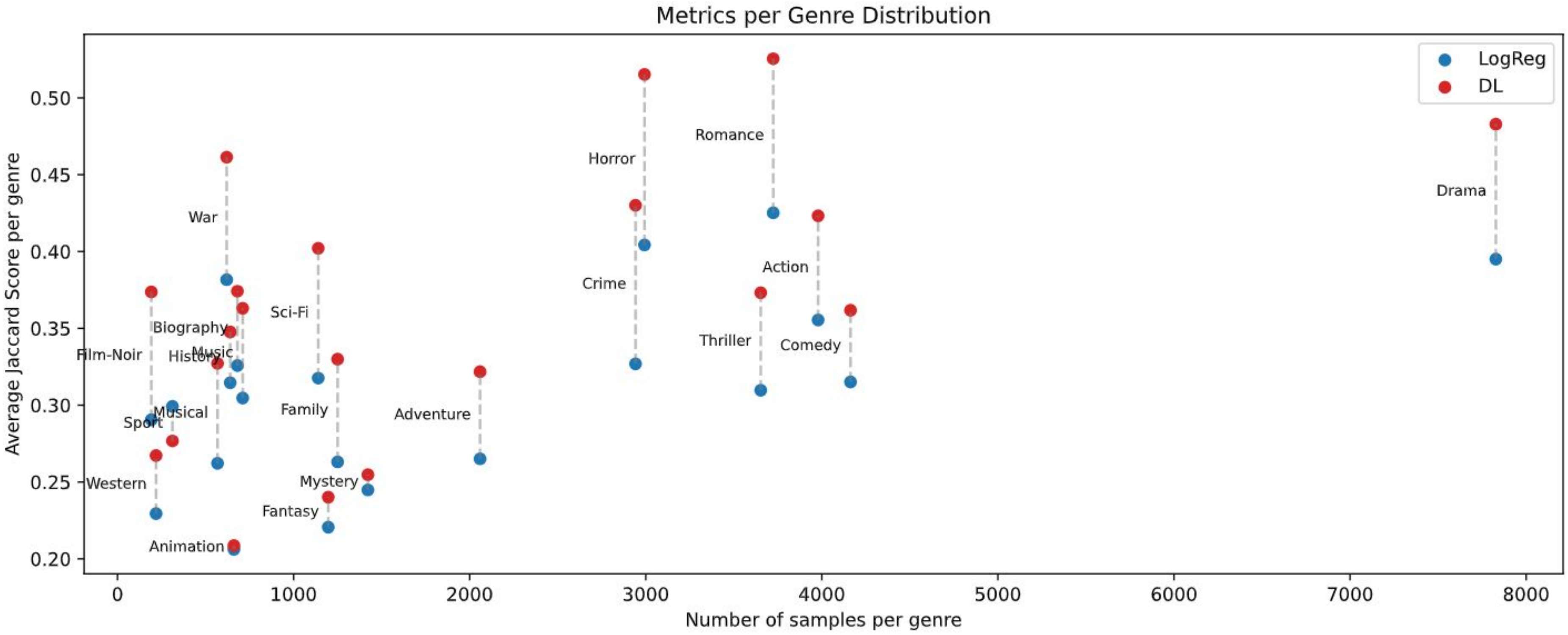
Metric	DL model	LogReg
Jaccard	0.42	0.38
Hamming Loss	0.09	0.10
Accuracy	0.14	----
Precision	**0.63**	**0.55**
Recall	**0.50**	**0.42**
At Least One	0.80	0.69
At Least Two	0.25	0.16

Focal loss

BCEWithLogitsLoss

class weights

DL vs. LogReg



Metrics per Genre Distribution
LogReg
DL
Average Jaccard Score per genre
Number of samples per genre
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0
1000
2000
3000
4000
5000
6000
7000
8000
War
Biography
Film-Noir
History
Music
Musical
Sport
Western
Animation
Sci-Fi
Family
Mystery
Fantasy
Adventure
Horror
Romance
Crime
Action
Thriller
Comedy
Drama

BREAK: SWITCH FROM DANIEL TO B

Identified Issues

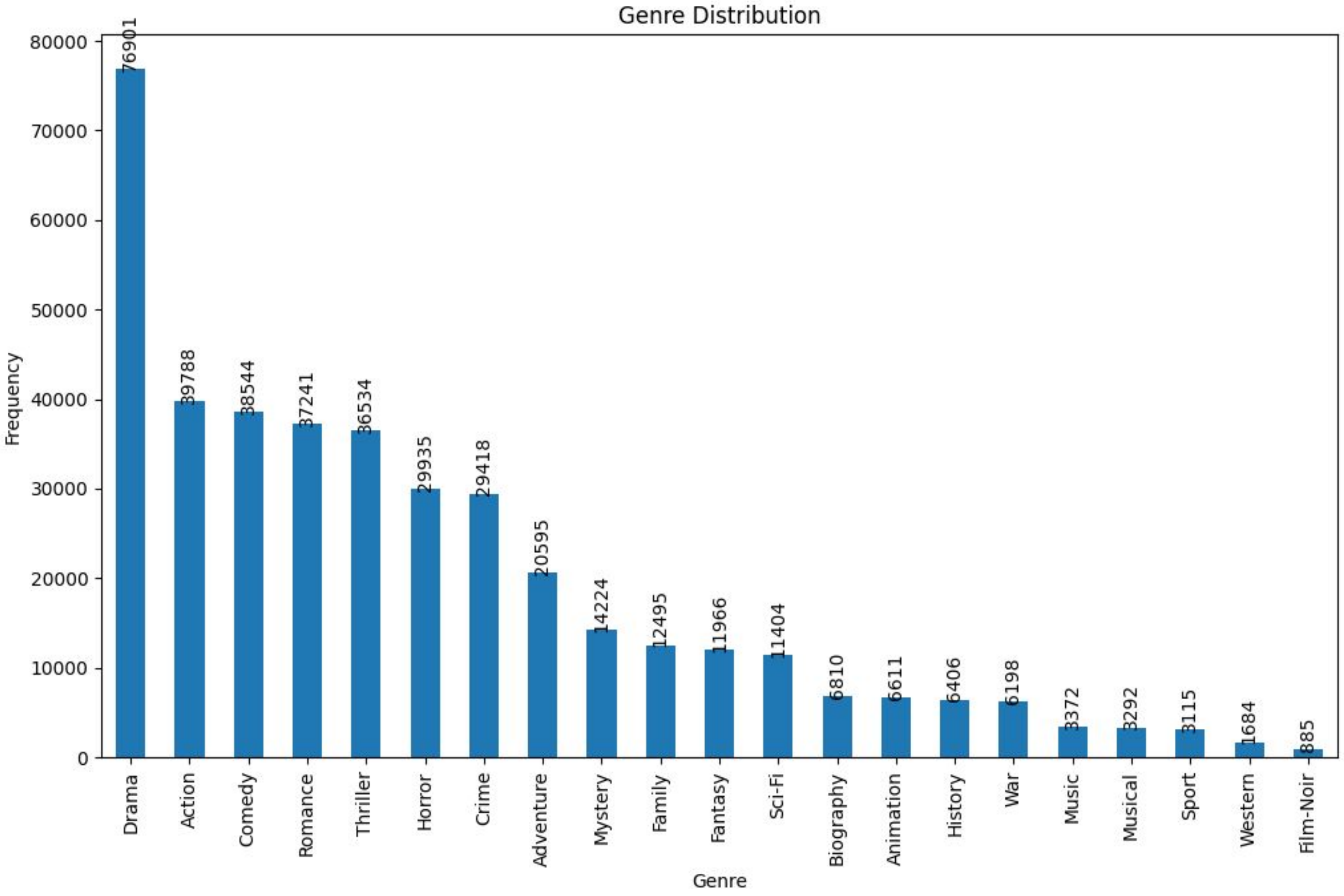
Issue - Movies without genre

There are Movies without genre???

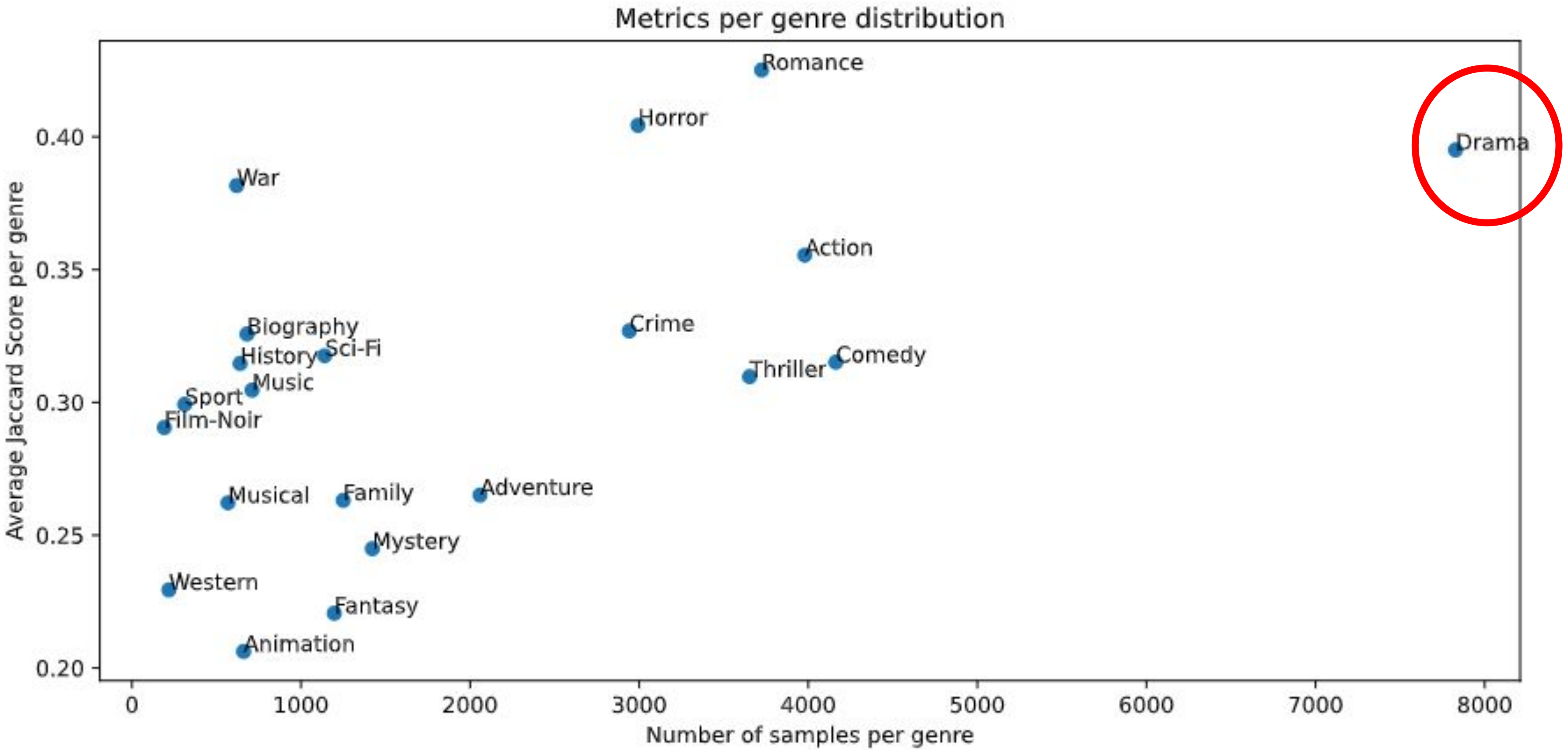
Predict-at-least-1 Mechanic

Force the MultiLabelClassifier to always predict at least one Genre, even if it has low confidence

Issue - Genre Imbalance



Metric / Genre-Distribution



If nothing then drama

"Trackhouse: Get Ready chronicles the launch of one of NASCAR's newest organizations."

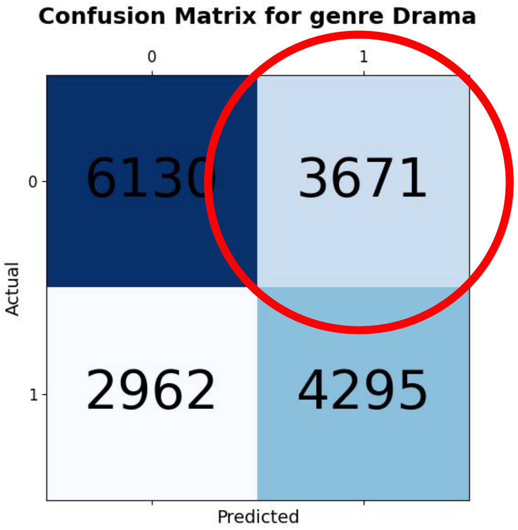
- **Labels**: ['*Sport*']
- **Predicted**: ['*Drama*']

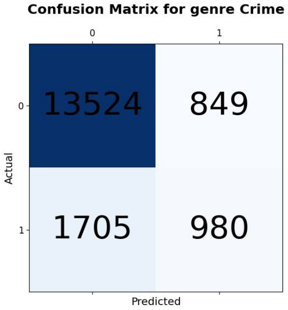
"The story of the highwayman and folk hero, Juraj Janosik."

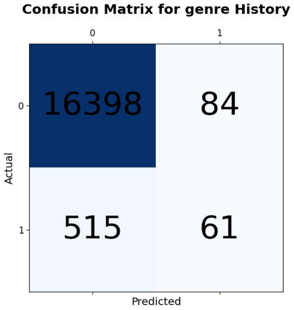
- **Labels**: ['*Animation*']
- **Predicted**: ['*Drama*']

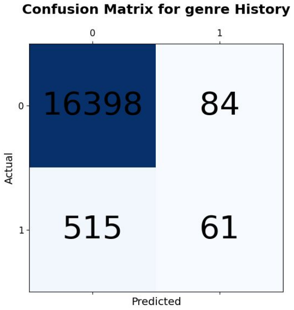
Confusion Matrix Drama

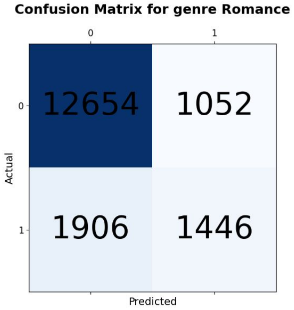
Many False Positives for Drama!





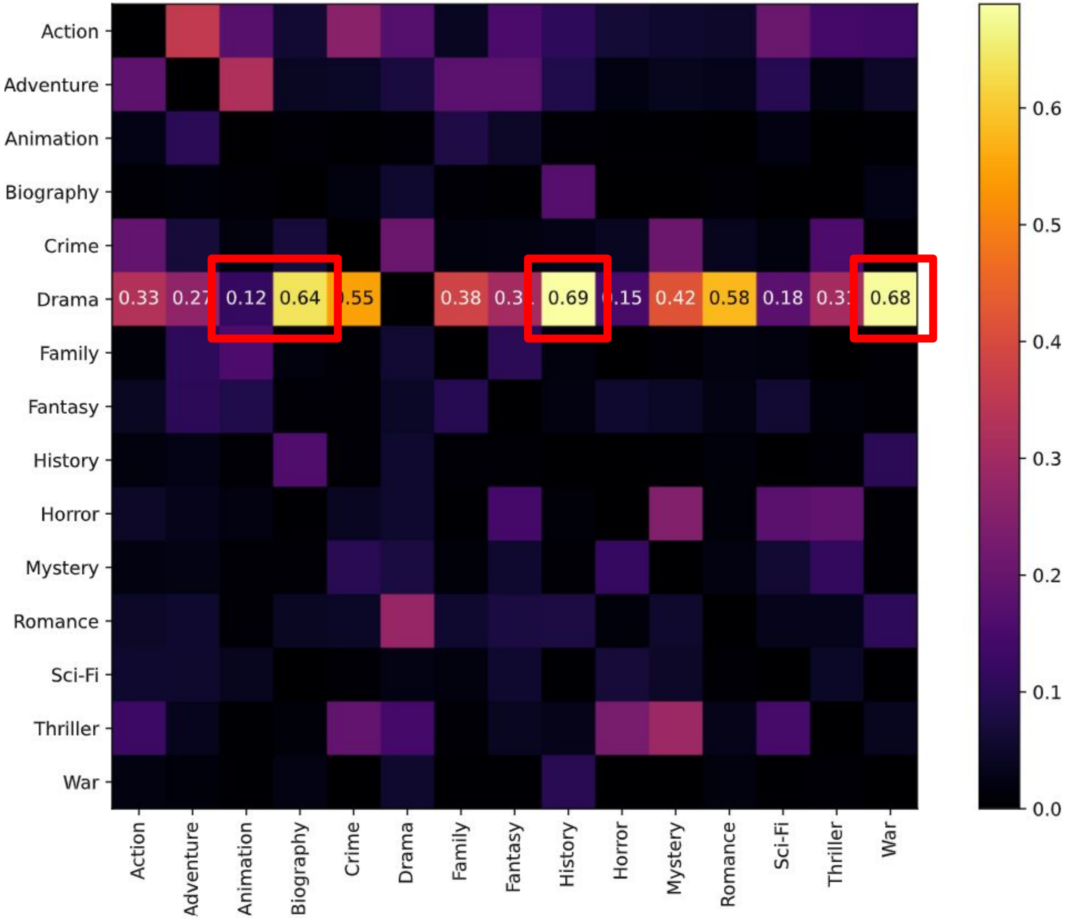






BREAK: SWITCH FROM B TO DANIEL

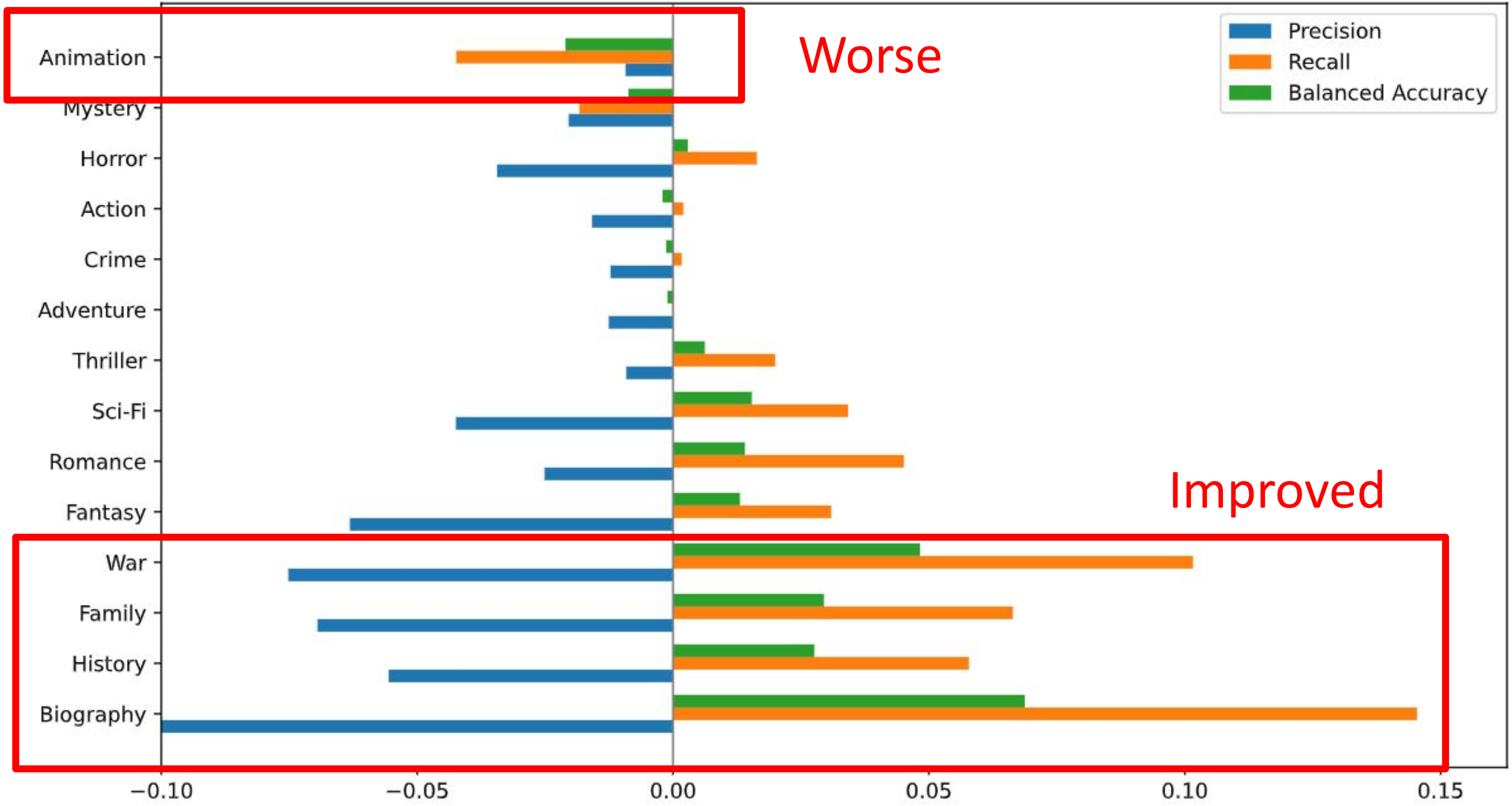
Genre co-occurrence matrix



History 0.69
War 0.68
Biography 0.64

Animation 0.12

Metric differences DL (NoDrama - Orig)



Feature Importances and Tf-idf vectors

[2.50, 2.11, 1.90, …. , -1.37, -1.43, -1.75]

normalisation

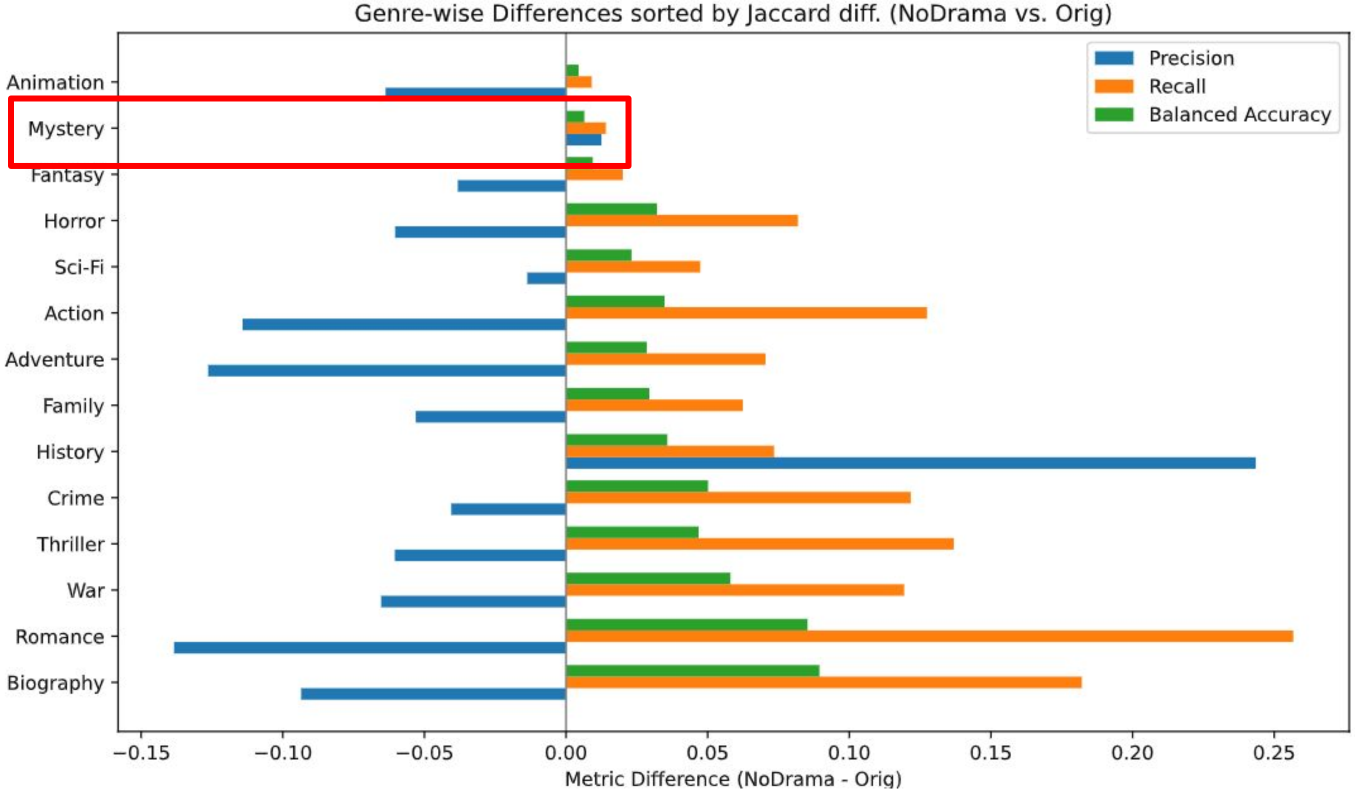
[1.00, 0.98, 0.96, …. , -0.92, -0.95, -1.00]

Feature Importances - Similarity to Drama

Genre	Cosine Similarity
War	0.36
Biography	0.35
History	0.32
Crime	0.30
…	…
Animation	-0.25
Horror	-0.52

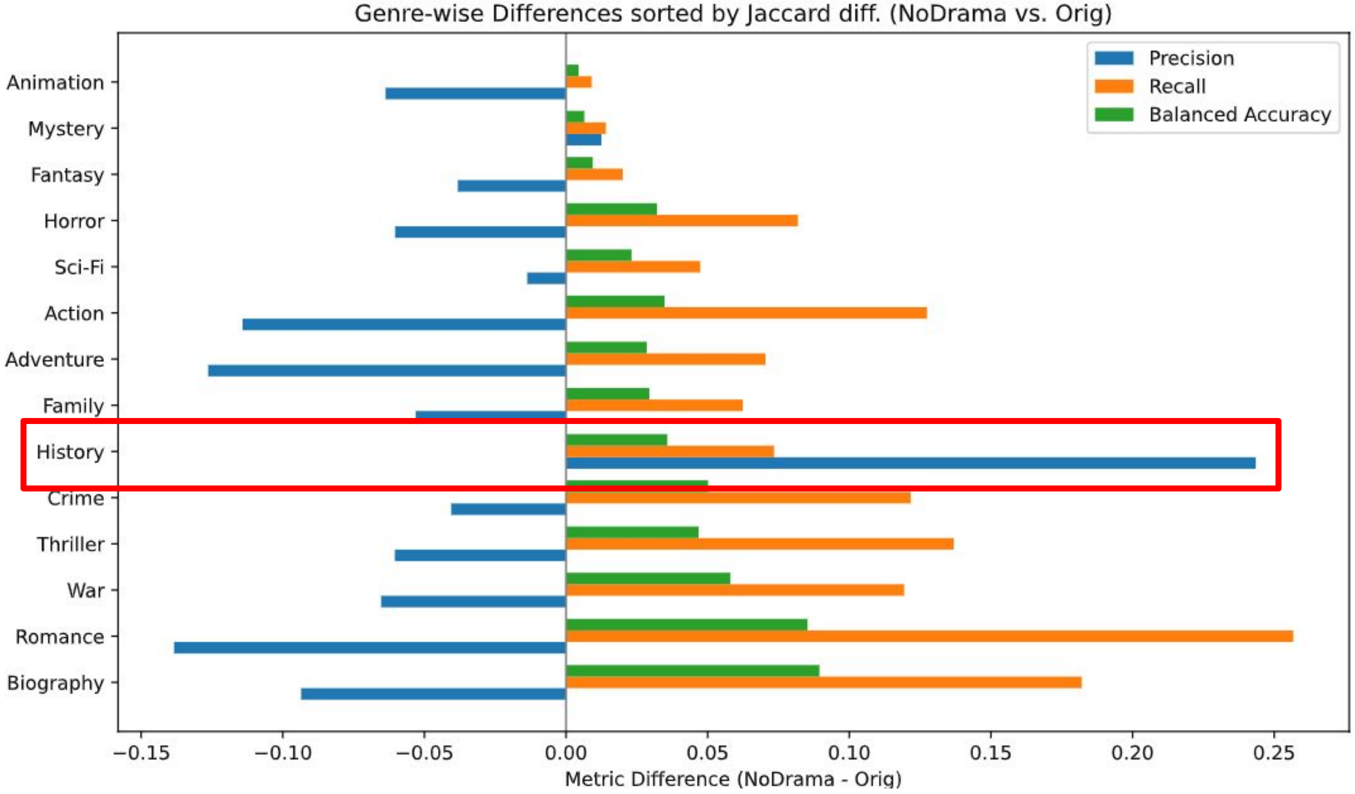
Genre	Euclidean Distance
Biography	6.957
War	7.432
Crime	7.834
History	8.485
…	…
Mystery	**10.528**
Horror	10.930

Feature Importances - Similarity to Drama



Genre	Euclidean Distance
Biography	6.957
War	7.432
Crime	7.834
History	8.485
…	…
Mystery	**10.528**
Horror	10.930

Tf-idf vectors - Similarity to Drama



Genre	Cosine Similarity
History	**0.8818**
Crime	0.8645
War	0.8452
Biography	0.809
…	…
Horror	0.6279
Animation	**0.5010**

Removing Drama - Changes in Predictions

→ Both co-occurrence with Drama and Support of genres matter
Low-support high co-occurrence see the biggest changes
Similarities of Feat. Importances and Tf-idf vectors confirms this

BUT!!!

BREAK: SWITCH FROM DANIEL TO C

Pruning Descriptions

We assume there are **different “types” of descriptions**.

Some actually describe content

- *“Anny works in a cigar shop. Wholesaler Willmann fancy Anny and hire her as her ….”*

Some are “Meta”-Descriptions

- *“The life of queen victoria”*
- *“An epic italian film, ‘Quo Vadis’ influenced many later works”*

Some contain author information at the end e.g.

- “... lawyer who has robbed him. [Synopsis from BIOSCOPE …]”

→ **Pruning Descriptions**

Oversampling

- MultiLabelClassifier trains **individual** Classifiers (e.g. LogReg)
- Ratio between positive and negative samples very unbalanced (much more negative then positive samples)

→ Oversampling

- **SMOTE** for Oversampling (generating new samples)
- Decided on a ratio of #pos_samples = 0.5 * #neg_samples

Experiments

- Predict at least one (!)
- Address class imbalance
- Pruning description
- Dropping “meta” descriptions (e.g. “starring, produced by, directed by”)
- Removing weird chars (e.g. “, “ “, -, “-,)
- Less Drama (?)
- Remove low-support genres
- Increase classifier probabilities based on frequently occurring words per genre (Hard-coded)

Results

CLF	Improvements	Jaccard	Hamming	Prec.	Recall	at-least-1
LReg	**Baseline**	0.333	**0.088**	0.521	0.385	0.623
LReg	**AL1**	0.377	**0.088**	**0.613**	0.428	0.700
LReg	**AL1,P**	0.378	**0.088**	0.608	0.430	0.716
LReg	**AL1,O**	0.378	0.096	0.554	0.487	0.753
LReg	**AL1,P,O,C**	**0.382**	0.116	0.487	**0.604**	**0.843**
DL	**-**	**0.423**	**0.086**	**0.621**	**0.504**	**0.785**

AL1=Predict-at-least-1 **O**=Oversampling **C**=Balanced Class Weights **P**=Prune

All LReg Models trained with lemmatized descriptions and tf-idf

Further Works

- **Include titles** and/or reviews of movies
- More sophisticated **description analysis**
- More sophisticated text modeling?
- Include information on Genre description
- Apply insights from removing Drama to the **predict_at_least_1** function

Future Works: Adding the Title

"Lance Hayward, a silent movie star, appears as various characters, killing quite a handful of unfortunates, using various weapons."

- **Labels**: [*'Horror'*]
- **Predicted**: [*'Action'*]

Note:

The title ("Terror Night") would provide the missing Horror signal.

Did you extract your information?

- Yes? Good
- No?

QUESTION!?