TU WIEN

Faculty of Infromatics

NLP
Movie Genre Classification
Management Summary

January 24, 2025

Daniel Chudý, Tim Dirr, Filip Faber, Sebastian Pinter

Overview

The goal of this project was to classify movie genres based on their plot descriptions using a multi-label, multi-class classification approach. We utilized a public accessibly dataset [1] and supplemental data from the IMDb API. We employed a simple linear model and a state of the art LLM to compare and evaluate the results.

Challenges

Dataset Quality: Over **120k** records out of the 190k total dataset entries contained incomplete or missing **plot descriptions**, which were replaced using the **IMDb API**. Also the quality and length of plot descriptions varies strongly.

Class Imbalance: The different genres are not evenly distributed, with Drama being the most common by far, resulting in bias in the predictions.

Genre Co-Occurance: Genres such as **Drama**, **Romance**, and **Thriller** very often co-occur, making it challenging for models to differentiate between them.

Resources

The **Kaggle-Dataset** mentioned in the overview is used as the foundational dataset. We call the **IMDb API** to substitute missing or incomplete **plots**. State-of-the-art **machine learning** and **NLP** Python libraries like **scikit-learn**, **stanza**, **nltk**, **PyTorch** and **Hugging Face Transformers** were utilized to build the models alongside a **pre-trained LLM** [2].

Solution

1. **Data Preprocessing:** The dataset was cleaned by removing irrelevant and nonsense descriptions in a very basic way by matching certain text patterns. Additionally, we removed very short descriptions as they may not contain valuable information. The dataset is stored in a state of the art format (CoNLL-U)
2. **Adress class imbalance:** Experiments with removing the **Drama** genre show that it helps to increase attention and recognition for **less-represented** genres. Additionally, we employed a data augmentation technique (SMOTE) to artificially generate new samples for underrepresented genres based on the prior existing plot descriptions.
3. **Model Development:** We used a basic logistic regression model and a more complex LLM to compare their results. The logistic regression model (LG) is easy to interpret where the LLM is better in extracting context from text. For the LG we employed basic feature extraction techniques based on counting word occurrence where the LLM is fed the raw text. We also force the models to always predict at least one genre as there must not be movies without a genre.

[1] `https://www.kaggle.com/datasets/rajugc/imdb-movies-dataset-based-on-genre?select=history.csv`

[2] `https://huggingface.co/distilbert/distilbert-base-uncased`

Insights

- **Drama's dominance** skewed predictions, often serving as a default genre when models were uncertain (caused by the `Predict_At_Least_One` logic).
- **Specific keywords** (e.g., ”future” for Sci-Fi) heavily influenced the predictions, creating an over-reliance on certain features.
- **Class imbalance** significantly impacted minority genres, as evidenced by a correlation between genre occurrence and performance scores.
- Also, there do not exist distinct definitions for certain genres and therefor making the assignment of genres to movies ambiguous and subjective even for humans.

Limitations

- **Ambiguous** or **very short** descriptions, that sometimes are ”nonsense” text.
- **Removing the Drama** genre **improved** predictions for other genres but **reduced** overall precision.
- More **computational power** needed to improve Deep learning model performance.

Future Work

Further ideas to improve the capabilities of the model could include the incorporate additional text from titles or/and reviews. Also, a more sophisticated filtering/preprocessing of descriptions could help with improving decision boundaries. As genre descriptions are very ambiguous, the labeling of the available data may not be uniformly accurate across different genres, as people may have different interpretations of genres. Re-Labeling with uniform genre definitions or including these definitions in the models could help to produce more meaningful predictions. Also if the models would predict one of the only-co-occurring genres, force them to do at least another prediction out of the set of genres the initial prediction always co-occurs with.

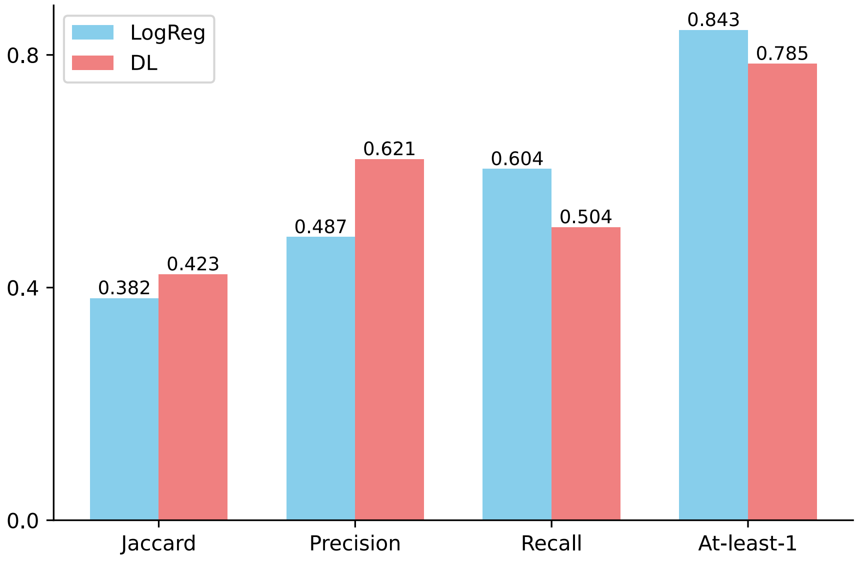

Figure 1: Classifer performance metrics.