Movie Genre Classification

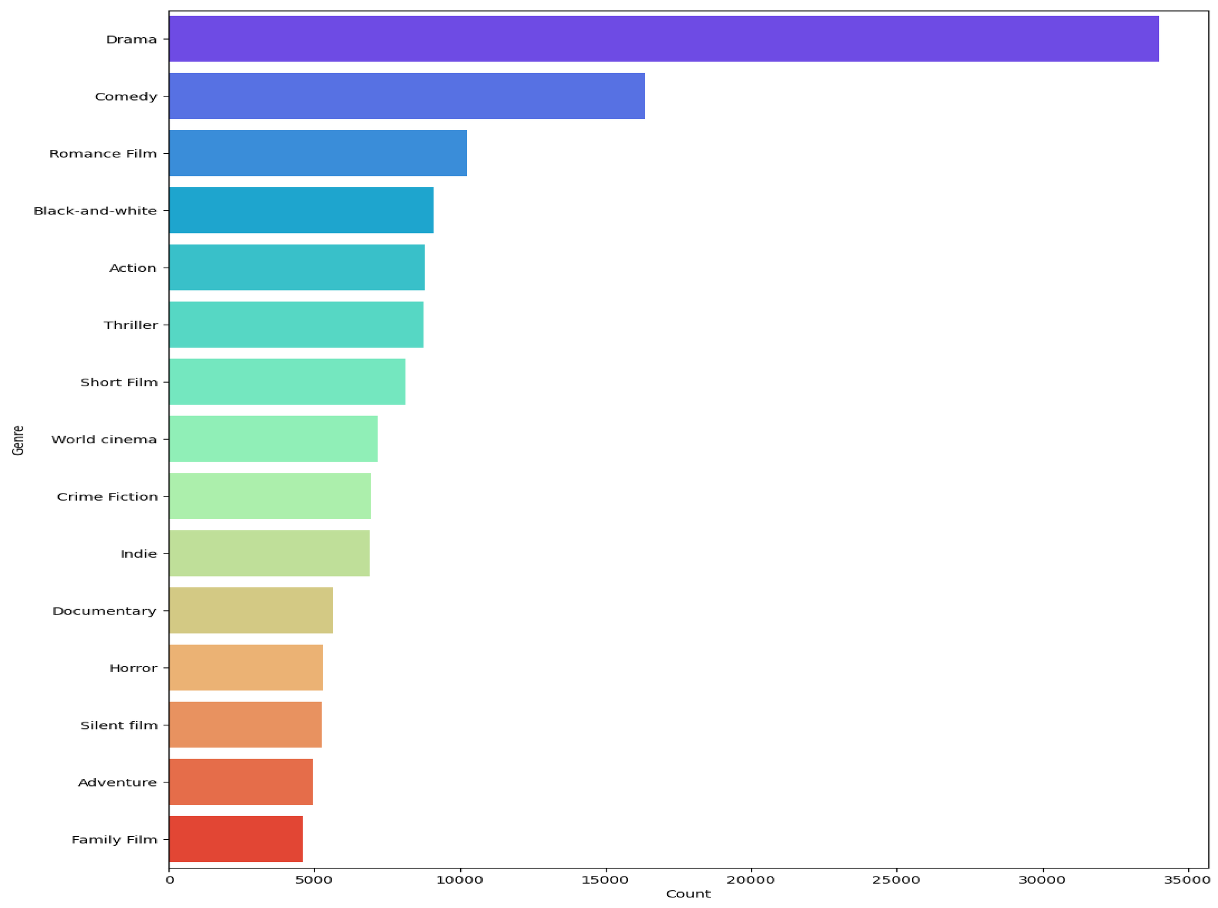
Amirhossein Soltani, Nigitha Selvarej

Introduction

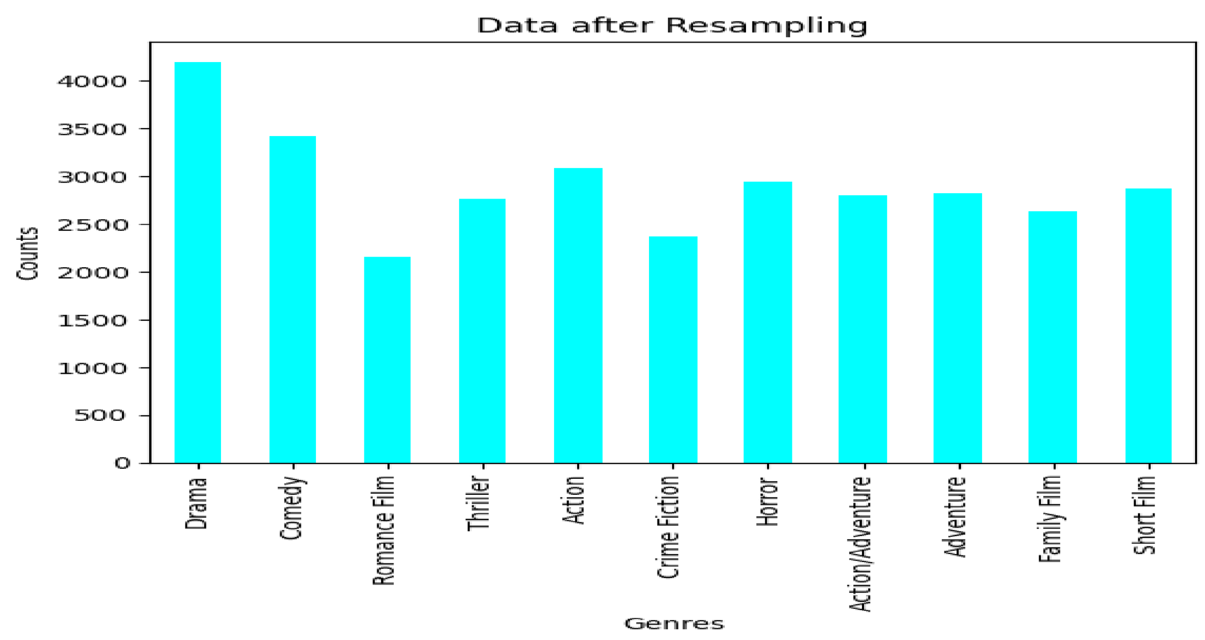
Project Overview:
Movie genre classification plays a crucial role in content recommendation systems by providing personalized movie suggestions, boosting user engagement. It significantly improves search functionality, allowing for efficient genre-based filtering and accurate search results.
This also provides valuable insights into audience preferences and market trends. This information is crucial for content producers and marketers to understand which genres are gaining popularity, which helps in making informed decisions about future content creation and marketing strategies.

Dataset:

- 42,306 movie plot summaries extracted from IMDB and Wikipedia, including 362 different genres, metadata about the actors and the movies.



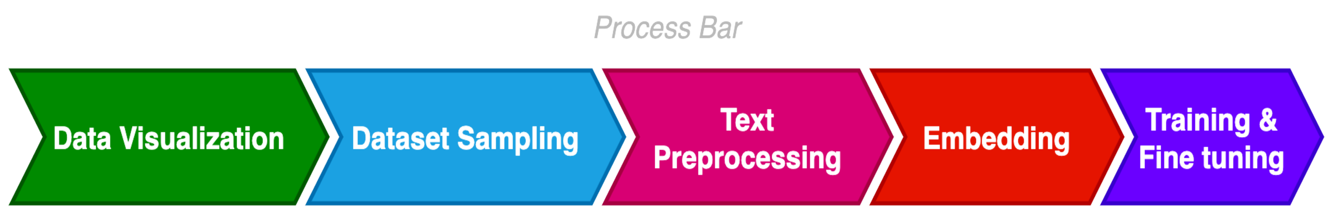
- Due to imbalanced number of labels for each movie genre, we needed to remove few-shot labels, and perform statistical sampling from the original dataset to prevent biased results from trained model.

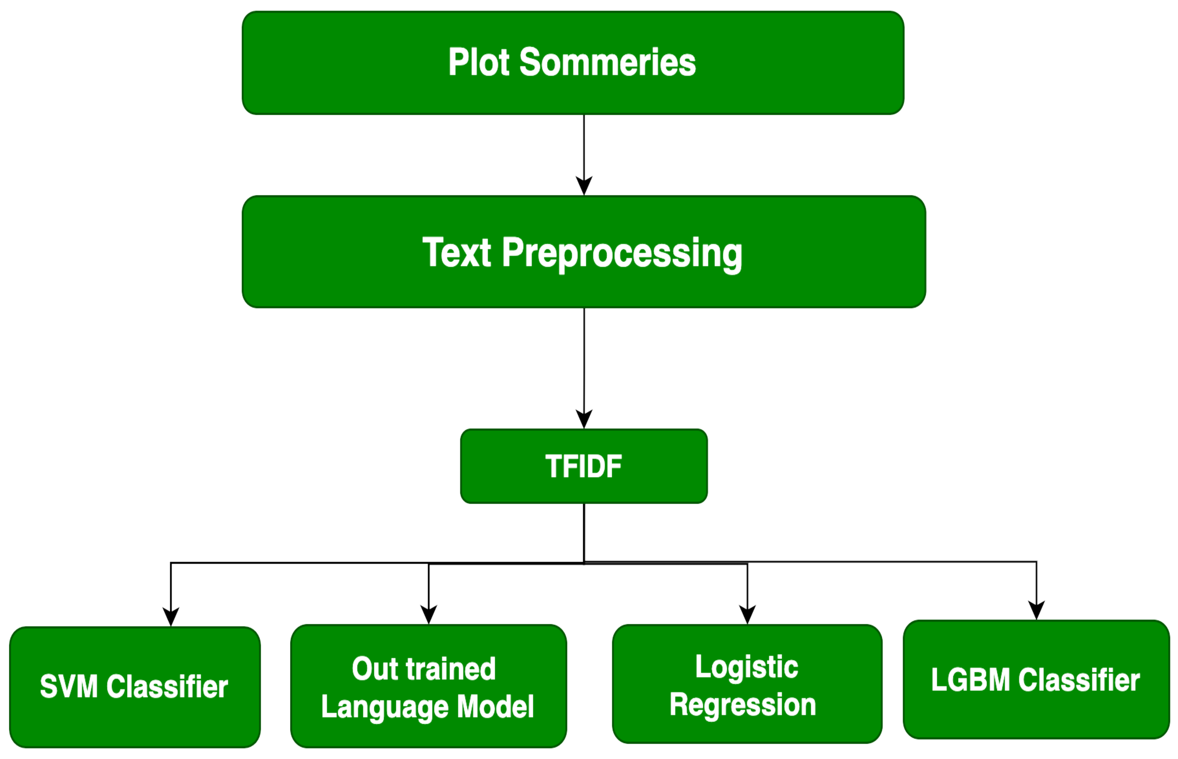


Methodology

In preprocessing step, we performed tokenization to convert text into tokens (words, sub-words, or characters) that the model can understand. Next steps included removing unnecessary characters, punctuation, or special symbols, normalization and lower casing, stemming and lemmatization.
Next, we performed vectorization to encode text data into numerical format to capture the contextual features of the plot summaries and make them suitable for machine learning models. We utilized TF-IDF with ngram size (1,3) to capture more relationships between words.

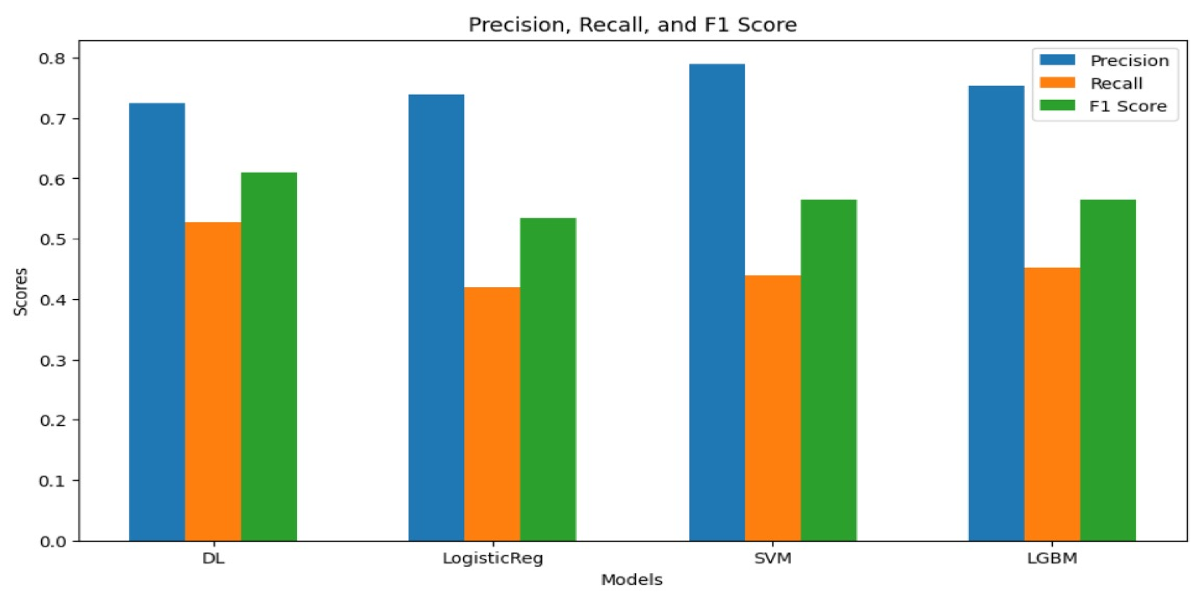
Finally, we trained our own language model as well as distinct classifiers and compared their results. Our language model is a Deep Learning (DL) model with multiple dense layers, batch normalization, regularization, and dropout for multilabel classification. It uses a Adam optimizer, step decay learning rate schedule, and early stopping for training the model.





Results

The DL model achieved the highest F1-score indicating superior performance in predicting movie genres. The SVM classifier has a higher precision but lower recall, indicating that it misses more instances of a particular genre as compared to the DL model.



Conclusion

Our model has a high precision indicating that the model's predictions for that genre are mostly correct, minimizing the number of false positives. The recall of the model is slightly lower but decent, missing only some true instances of a particular genre. The model's validation performance is close to the training performance indicating good generalization capability.
Future work includes experimenting with different feature set and hyper-parameter tuning to optimize the model performance. More advanced models could be explored to understand the sequence and context of texts.

References

[1] Yinglong Ma, Xiaofeng Liu, Lijiao Zhao, Yue Liang, Peng Zhang, and Beihong Jin. Hybrid embedding-based text representation for hierarchical multi-label text classification. *Expert Systems with Applications*, 187:115905, 2022.

[2] Linkun Cai, Yu Song, Tao Liu, and Kunli Zhang. A hybrid bert model that incorporates label semantics via adjustive attention for multi-label text classification. *IEEE Access*, 8:152183–152192, 2020.

Contact

Nigitha Selvraj
Hochschule Bonn-Rhein-Sieg
Email: nigitha.selvaraj@smail.inf.h-brs.de

Amirhossein Soltani
Hochschule Bonn-Rhein-Sieg
Email: amirhossein.soltani@smail.inf.h-brs.de

b-it Bonn-Aachen International Center for Information Technology

Hochschule Bonn-Rhein-Sieg University of Applied Sciences