Problem Identification & Project Specification

Farida Bey
900212071

&

Youssef Ghaleb
900211976

INTRODUCTION

Decisions are made and supported by fact, preference, or both. Every part of daily life requires us to make decisions, especially when it comes to entertainment. The entertainment industry is filled with content that caters to individuals from all walks of life, each with their own preferences. The film industry specifically provides watchers with a large range of themes and genres to choose from. Each individual prefers a specific genre over another, which is what most streaming platforms analyze and utilize to keep user engagement high. As the volume of movies offered on the several streaming platforms available to consumers increases, movie genre classification becomes a more essential function of these platforms. Allowing for efficient content management, organization, and recommendation to maximize user experience.

Motivation:

In today's world, it is unrealistic to manually classify which movies fall under a specific genre. The utilization of machine learning provides a more convenient, efficient, and scalable approach for movie genre classification. Achieved through the use of different training algorithms and metadata.

Problem specifications:

Movies are complex, often with many underlying themes, they conform to several genres. However, the goal is to develop a machine learning model that accurately specifies which genre a movie falls under most. Thus, signifying our problem, to be a classification problem.

LITERATURE REVIEW

I. Article 1:

In their study, the authors introduced CNN-MoTion, a specialized Convolution Neural Network architecture for movie trailer genre classification. They developed a novel dataset of over 3500 trailers with known genres and compared CNN-MoTion against traditional feature extraction methods like Gist, CENTRIST, w-CENTRIST, and low-level feature extraction techniques. The results showed that CNN-MoTion significantly outperformed these methods, highlighting the CNN's effectiveness in capturing complex patterns and features essential for accurate genre classification [2].

II. Article 2:

The paper investigates the efficacy of using movie posters and overviews to predict movie genres by treating it as a multi-label classification problem. The authors propose that combining visual features from posters with textual features from movie overviews provides better genre prediction than using each type of feature individually. They utilize a ResNet50 model for visual features and GloVe embeddings for textual features, achieving an average F1-Score of 0.674, which indicates improved performance over traditional models. The study emphasizes the importance of both posters and overviews in conveying genre information and highlights how their combined use enhances predictive accuracy [5].

III. Book 1:

The book discusses a method for classifying movie genres from posters using multi-label classification techniques. It emphasizes that movie posters contain significant visual cues that can aid in genre detection, making this task suitable for automated approaches. The study employs three classification methods—ML-kNN, RAKEL, and Naïve Bayes—applied to a dataset of 6,739 movie posters labeled with up to 18 genres. The research focuses on using low-level visual features like color histograms, dominant colors, and GIST descriptors. Results showed that Naïve Bayes achieved the highest F1 score of about 0.38 when using the complete feature set for 11 genres. The study highlights that while feature reduction did not significantly affect performance, dominant colors were crucial for genre classification. Future work includes testing additional features and larger datasets and comparing automatic results with human assessments [1].

IV. Article 3:

The study investigates movie genre classification from plot summaries using bidirectional LSTM (Bi-LSTM) networks. It involves dividing each plot summary into sentences, assigning genres to each sentence, and then training a Bi-LSTM model with these sentence-level representations. The final genre prediction is made through majority voting based on the sentence-level predictions. The results show that this sentence-based approach with Bi-LSTM outperforms methods that use the entire plot summary at once and also performs better than basic Recurrent Neural Networks (RNNs) and Logistic Regression. This approach enhances the ability to capture genre-specific information from plot summaries, which can be useful for movie recommendations and validating plot accuracy [3].

PROJECT OVERVIEW

While the previously discussed projects offered invaluable insights, many of their approaches (such as the use of different neural networks and deep learning algorithms) are considered to be out of scope for this project. We have drawn inspiration from them and plan on utilizing natural language processing to analyze movie titles as well as movie descriptions. While also giving the features of our data set different weights as features such as IMDb rating, director, actors or runtime are not defining traits of a specific genre. It is for that reason that features such as movie titles and descriptions will be given more weight. In addition, one hot encoding will be applied to convert the movie genres into binary format so that the model can be trained to predict the genres as categorical labels. Moreover, our research has led us to believe that the best models to use are the random forest model and decision trees due to their compatibility with data processed through NLP methods.

This model can be applied to better recommend movies based on different user inputs such as; recommending a movie based on a user's preferred genre, or taking a movie title and description as input and recommending a movie of that same genre. In addition, the model can be used for content organization on streaming services. Moreover, the model could be used as a tool to study trends in movie genres and watchers preferences over time. Offering companies such as Disney, Fox, and Colombia pictures unique insights that will allow them to make informed decisions and capitalize on the information learned.

Bringing us to the completion of the project's first phase which involved: finding the data, identifying the problem, and specifying the scope of the project. Which has been discussed above. The second phase of the project will require us to filter through the data, deleting any rows with empty or invalid data to maximize the efficiency of the model. In addition, the data set we have selected contains multiple csv files storing feature data per genre. Completing the process which is known as feature engineering. We then need to compile and organize all the data from the different csv files into one file for efficient and effective training and testing. After which the file will be split into two data sets, training and testing. Allowing for phase three to begin where an exploration and study of different supervised learning models will be done on our data to assess their respective benefits and weaknesses within the context of our project. We will also be sure to explore other models if we find our previously planned decision tree and random forest models to be insufficient. Finally allowing for the implementation, training and testing of the final model concluding the fourth phase of the project.

DATASET OVERVIEW

While searching for datasets we found multiple datasets that could work and we chose one that we think is the best fit for our project.

1. **IMDb Movie Dataset: All Movies by Genre**

   Source: IMDb Movie Dataset on Kaggle

   This dataset provides detailed information about movies including but not limited to title, genre, description, director, and stars. It is suitable for genre classification tasks and aligns with your project goals. The dataset is divided into files based on the genre, each file contains from 15k to 50k instances and there are 16 genres, so we are expecting *approximately 400k instances*. The dataset has *14 features*.

   We found this dataset relevant to our project as it has a large number of instances that are big enough to train our model. It also has all the key features we're focusing on and it is classified by genre which is our label. The description field is especially useful as it will allow the model to capture more nuanced relationships between plot themes and genres, and the cast/director details add further depth for classification.

While this dataset is extensive and includes valuable information, depending on the completeness of the data it might need to be cleaned to remove the rows with null cells, and we will also need to merge all the files into one table as it is divided by genre. The main limitation is that the dataset captures multiple genres per movie, so we will need to select the most prominent genre for each film. This is our chosen one.

2. **TMDB All Movies Dataset**

Source: TMDB All Movies on Kaggle

This dataset includes movies from the TMDB (The Movie Database) API and also provides detailed information about movies like the above, title, genre, language, overview, etc. It is suitable for genre classification tasks as it has more features than the above with *26 features* and about *600k instances*.

We found this dataset relevant to our project as it has a large number of instances that are big enough to train our model. It also has all the key features we're focusing on. The overview field is the same as the description in the previous dataset, which will also require text processing.

While this dataset is extensive and includes valuable information, it mixes movies and TV shows, which may require filtering to meet our project's focus on movies, and because of its large size, it will take a lot of time to filter the data. Moreover, depending on the completeness of the data it might need cleaning, and the dataset captures multiple genres per movie, so we will need to select the most prominent genre for each film.

3. **Genre Classification Dataset IMDb**

Source: Genre Classification Dataset IMDb on Kaggle

This dataset is designed for movie genre classification, including features like movie titles, descriptions, and genres, having *10k instances and 3 features*. Thus it is not suitable for our project as it is very small for training our model, and it has limited features.

4. **Movie Genre from its Poster**

Source: Movie Genre from its Poster on Kaggle

This dataset includes movie posters along with metadata like titles, genres, and release years. It has around *40k instances* and *6 features*. It also has the movie poster as one of the features; however, it has the links to the posters and not the actual posters themselves. This dataset is out of the scope of our project as it will require the use of neural networks. Even if we ignored the posters we would be left with a very small dataset not suitable for training.

In conclusion, while we were initially hesitant between the first two datasets due to their size and extensive features, we ultimately chose the IMDb Movie Dataset because it aligns more closely with our project's goals and requires less extensive cleaning compared to the TMDB dataset.

REFERENCES

[1] Marina Ivasic-Kos, Miran Pobar, and Ivo Ipsic. 2015. Automatic Movie Posters Classification into Genres. In *Advances in intelligent systems and computing*. 319–328. https://doi.org/10.1007/978-3-319-09879-1_32

[2] 2016. Movie genre classification with Convolutional Neural Networks. *IEEE Conference Publication | IEEE Xplore*. Retrieved from https://ieeexplore.ieee.org/abstract/document/7727207?casa_token=k5QXmkYAkecAAAAA:ecEptJOvm2VCpEaKSvjevy7S6PkT0UBMVfUX72uqoxRIa3d8xyeS9Vc1Wk_o9YnQXNi1cy9J

[3] 2018. Movie Genre Classification from Plot Summaries Using Bidirectional LSTM. *IEEE Conference Publication | IEEE Xplore*. Retrieved from https://ieeexplore.ieee.org/abstract/document/8334466/

[4] 2018. Movie Genre from its Poster. *Kaggle*. Retrieved from https://www.kaggle.com/datasets/neha1703/movie-genre-from-its-poster

[5] 2020. Multi modal genre classification of movies. *IEEE Conference Publication | IEEE Xplore*. Retrieved from https://ieeexplore.ieee.org/abstract/document/9298385?casa_token=X_A5vudXB_8AAAAA:k2C7w65PpUboRoHVab42xlyHwKXYCIqcYMcshHxOoWF-OfsOaR-wJE5TwlLrzmERD73myA9H

[6] 2021. tmdb 650234 all movies and TV Shows. *Kaggle*. Retrieved from https://www.kaggle.com/datasets/mathlasker/tmdb-allmovies

[7] 2021. Genre Classification Dataset IMDB. *Kaggle*. Retrieved from https://www.kaggle.com/datasets/hijest/genre-classification-dataset-imdb

[8] 2023. IMDB Movie Dataset: All movies by genre. *Kaggle*. Retrieved from https://www.kaggle.com/datasets/rajugc/imdb-movies-dataset-based-on-genre