IMDB Movie Analysis Recommendation Along with Computer Vision

Mr. Chethan Reddy C V[1], Ms. K Sukanya[2], Mr. K Hari Charan[3], Ms. R Chidvi[4], Mr. Narayana H M[5]
[1,2,3,4]*Student,* [5]*Associate Professor, Department of CSE, M. S. Engineering College, Bangalore, India*

Abstract The showcasing entertainment industry is known as the most powerful andimpactful industries in the world. We all consume movies, tv shows, docuseries all of us are contributors to the global success of the industry. As big as it is, it also creates different feelings and opinions on the public. Some people love a certain movie, while others may dislike it. It is what it is, we all have differenttastes, we all have our preferences and that is why the industry must keep reinvent itself. Actors and actresses need to constantly push themselves to be more versatile, writers and producers must study ways of improving the quality of their movies. it is a process in constant development. Nowadays, people are more demanding than ever. With the insane amount of movie production on the market, it is also very easy for us, consumers,to jump from a movie to another, from a series to another. So, the decision factors for anyof us to choose a certain movie to watch are many times in the details. Details maybe howexciting a trailer is, if it was shot in a certain country, the sound effects on the trailer, which company produced it, which actors and actresses are in… (Computer vision can be a greattool to help us understand the importance of this features) Through the IMDb data base, it is possible to develop analysis and understand what might influence the ratings. Regression models, convolutional neural network and keras models are examples of possible paths in order to explain the dependent variable (movie rating).

I.INTRODUCTION

This project aims at analysing movie data – available at the IMDb database – explore its evolution over the years concerning different features, develop a model to predict ratings and increase our knowledge in computer vision through a small example of its use. Living through a pandemic is an unprecedented situation (in the most recent decades) that surely took a tool on the entertainment industry. Further on the project, it will be possible to visualize the evolution of the number of movies produced by year. Anintense decrease is expected when comparing 2020 to previous years. When focusing on movie ratings, one big question that arises is: which features are determinant to the rating? In fact, many characteristics can lead to a movie having a higher or lower evaluation. For example, is the movie genre relevant to the overall rating?Can certain actors and actresses increase (or decrease) the movie rating? To be able to come up with answers to these questions, data will be explored and a sample will be usedto develop prediction models – having as regressors the previous features. Lastly, we want to push our boundaries when it comes to computer vision and develop a small model that will get the movie name from its front cover and automaticallysearch it on the IMDb data base. With the advancement of information technology and the easier access to the internet, people can easily access various information about movies. There are many websites that provide information about movies such as Internet Movie Database (IMDb), Rotten Tomatoes, Metacritic and The Movie Database (IMDb). These websites provide information about movies such as actors, directors, budgets, ratings and user comments. Among these websites, IMDb is the best consumer website that contains information about movies, such as: financial information, ratings, casts, reviews, crew, actors, directors, summaries, story lines etc. This site contains a large amount of data, which contains a lot of valuable information about general trends in movies. An accurate movie rating prediction can help people to determine which movie to be watched. In addition, rating predictions are also beneficial for the economy. User ratings is form of Word of Mouth (WOM), rating which is as statement made by consumers about a product that is available to other consumers on the internet. User ratings are a good indicator for predicting the sales performance of a product in the future. Film industry experts agree that

rating is a key factor in film success and helps film production companies and investors to gain financial success. Companies can see which movies are likely to have good rating sand make strategies to utilize the movie to increase profits such as, by making merchandise such as, making merchandise for movies or create events and promotions related to the movie. Additionally, by utilizing the historical values obtained from previously released movies, rating predictions can be made before the film is produced. Film maker companies can make strategic plans and decisions.

II. EXISTING SYSTEM

The first (Augustine, A., & Pathak, M. (2020)) they develop a model prediction based on Neural Networks mainly based on crew members information. It was important to go through this research since the accuracy obtained is fairly low – which also shows us the importance of testing different models and try different paths to achieve the best possible results. Using one dimensional convolutional neural network from (Abarja, R. A., & Wibowo, A. (2020)) proved to be an accurate approach to predict movie ratings when working with somehow small samples. On the paper (Dixit, P., Hussain, S., & Singh, G (2020)), the authors explore regression and classification models, trying different approaches and obtaining the best accuracy when using the Gradient Boosting classification model.

III. PROPOSED SYSTEM

This study used datasets obtained from Kaggle. The datasets have comma separated value (CSV) format. There are two datasets, the first dataset source contains movie data from IMDb and the other contains movie data from IMDb. These datasets were combined into one dataset based on movie title and release year. After the dataset is collected, pre-processing process such as data cleaning, data transformation, and feature extraction are executed. The creation and selection of features aim to provide a good quality dataset that can improve the performance of the movie rating prediction. Features are created from movie attributes that have relationship with the success of a movie. The features are categorized into the following categories: historical features, numerical features, categorical features, topical features, and social media features.

Features

Features are created from movie attributes that have relationship with the success of a movie. The features are categorized into the following categories: historical features, numerical features, categorical features, topical features, and social media features.

IV. SYSTEM ARCHITECTURE

System architecture refers to the fundamental structure and organization of a complex system, such as a computer system, software application, or an entire information technology infrastructure. It involves designing and describing the various components of the system, their relationships, and how they work together to achieve the desired functionality and goals. System architecture encompasses both the high-level conceptual design and the detailed technical specifications of a system. It defines the system's overall structure, the arrangement and interconnections of its modules or components, the data flows, communication protocols, and the interfaces between different subsystems or external entities.

The purpose of system architecture is to provide a blueprint or roadmap for building and maintaining the system. It helps in ensuring that the system meets the required performance, scalability, reliability, security, and other quality attributes. System architecture also enables system designers, developers, and stakeholders to have a common understanding of the system and facilitates communication and collaboration among them.

System Architecture is an organized description that defines the structure, behaviour, and the system views. The system architecture describes the major components, their relationships, structures and how they interact with each other. Software architecture and design includes several contributory factors such as Business strategy, quality attributes and many more. Software Architecture and Design can be classified into two phases - Software Architecture and Software Design. Software architecture refers to the fundamental structure of a software system where each structure comprises of elements of the software, the relationship between them and the properties of elements and the relations.

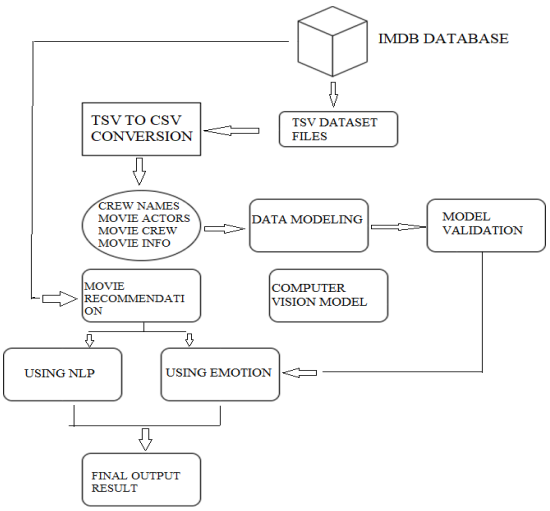

Fig 4.1 System Architecture

4.METHODOLOGY

The methodology followed is the CRIST-DM/POST-DS: business understanding, dataunderstanding, data preparation, modelling and model validation.

Business Understanding

In IMDb website, we can find several datasets, regarding information about all kinds of tv shows and movies and its crew. Analysing this type of information can be a powerful tool to understand what the consumer is looking for a movie or a tv show. When writers or directors plan their movies, they can have a better understanding ofwhich actors and genres work better to achieve the best rating possible.

Data understanding:

To develop this project, we worked with 5 IMDb data sets, regarding movies from2010to 2021:

- Basics Title

Data set containing a row per movie/series, movie/series code as index andcolumns:

Title Type : type of title (ex: for movies, 'titleType': short; for episodes oftvshows, 'titleType': tvEpisode)

- Primary Title : movie title;
- Original Title : original movie title;
- Is Adult : 0 if movie is not Adult; 1 if it is;
- StartYear : release year;
- End Year : year in which the series stopped being streamed. '\N' for allmovies;
- runtime Minutes : movie duration, in minutes;

Note that this data set also includes data regarding series, not only movies, but ‘onlymovies'data were considered for this project.

- Title Ratings

Data set regarding movies' ratings. For each movie, stores by column:

- Average Rating: stores the average rating attributed to the movie (on a scale from 1 to 10);
- Num Votes: number of votes on the corresponding movie;

- Title Crew

Data set with crew information (directors and writers) per movie. Each individuals given a code, so there are no names on this data set. Columns:

- Directors : codes of the directors of the movie;
- Num Votes : codes of the writers of the movie;

On both columns, each cell can store more than one code, doing so separating codes by commas. (ex: code1, code2)

- Title Principals

Data set storing information regarding all professionals who took part on the movie Columns:

- T const : movie code;
- Ordering : works as an individual index for each movie. Integers starting at 1, incrementing 1 by each row while the movie code does not change. When the entire movie crew is characterized, 'tconst' column receives a new code and 'ordering' starts from 1 again;
- N const : crew member code;
- Category : stores the in which category the person works on this particularmovie;
- director of photography);

- Name Basics

Data set storing information for each professional related to movies. Has as columns:

- Primary Name : person's name;
- Birth Year : Year in which he/she was born;
- Death Year : Year in which he/she died. If the corresponding professionalis alive, stores '\N';
- Primary Profession : profession(s) related to movies;
- Known For Titles : Movie codes in which he/she worked on. Can storemultiple codes, separated by commas (ex: code1,code2,code3).

Data preparation:

- Data preparation for visualization
  - Filter the dataset on movies from 2010 to 2021.
  - Delete all the rows with NA values on start year and genre.
  - Merge the rating data frame to the Title Basics data frame.
  - Filter all the other data frames on the movies that are on the Title Basics data frame.
  - Create dummies on the Title Basics data frame to all genres.
  - Export all this new table to csv files.

- Data preparation for rating prediction
  - Filter the visualization dataset on movies from 2020 to 2021.
  - Remove all movies with NA values on runtime Minutes.

Modelling and validation

To model this dataset, we selected a set of models: OLS, Ridge (alpha = 0.5), Lasso (alpha = 0.5), Bayesian Ridge, Neural Network and a Sequential Model from keras.We then measured the quality of the models using 3 indicators: R^2, Mean Absolute Error and Mean Squared Error. After developing the model and choosing the best one, each movie can have a predicted rate. Having this prediction writers and directors can:

- Chose the optimal crew and movie genre.
- Maximize the movie rating.
- Understand what the audience prefers.

V. IMPLEMENTATION

By choosing the year, the user will get the average rating for each movie genre (for movies produced on the corresponding year). For simplification, on this report, we present the top four genres from 2019 and 2020.Contrary to what happened regarding the number of movies produced by genre, the top 4 genres of highest average movie rating change from 2019 to 2020. Note that theratings attributed to the movies are in a scale from 1 to 10 (10 being the best possible evaluation and 1 the worse). One thing to note is the consistency of Documentary movies:not only it is one of the preferred genres in both 2019 and 2020, but also one of the most type of movies produced. This tells us that Documentaries, in general, are both intensively produced and of very high quality. Finally, also note that Talk Shows reached the first position in 2019 with a 10.0 average rating, and only fell down to second place in 2020, only 0.062 behind the highestrated genre (documentaries). Natural Language Processing (NLP) is a subfield of artificial intelligence (AI) that focuses on the interaction between computers and human language. It involves the development of algorithms and models that enable computers to understand, interpret, and generate human language in a way that is meaningful and useful.

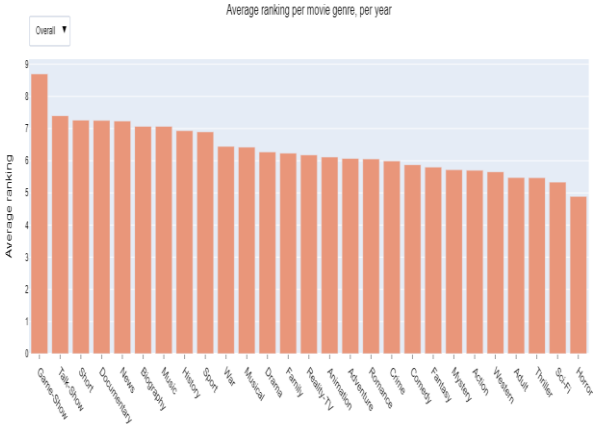

Fig 5.1

VI. RESULTS

the project has successfully developed a system that provides accurate and relevant movie recommendations to users. Through the analysis of movie attributes, such as genres, release years, and ratings, valuable insights have been gained regarding the distribution and trends within the movie industry.

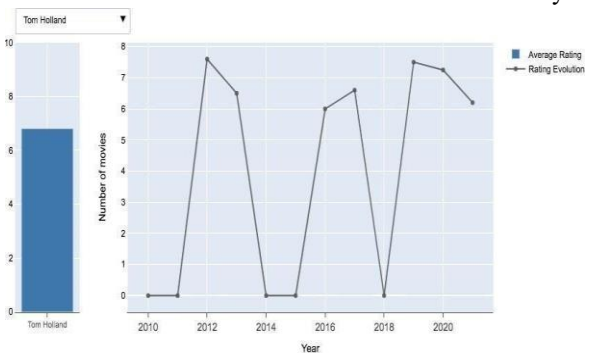

Fig 6.1

VII. CONCLUSION

The project on IMDB movie analysis and recommendation, coupled with computer vision integration, has addressed the challenges posed by

information overload and the need for personalized movie discovery in the entertainment industry. By leveraging data analysis techniques and recommendation algorithms, the project has successfully developed a system that provides accurate and relevant movie recommendations to users. Through the analysis of movie attributes, such as genres, release years, and ratings, valuable insights have been gained regarding the distribution and trends within the movie industry. These insights have enabled the creation of a recommendation system that takes into account user preferences and historical data to generate personalized movie recommendations. Furthermore, the integration of computer vision techniques, specifically in analysing movie posters or stills, has enhanced the recommendation system's accuracy and relevance. The extraction of visual features from images has enabled a deeper understanding of the visual aspects of movies and their correlation with user preferences.

REFERENCE

[1] Augustine, A., & Pathak, M. (2008). User rating prediction for movies. Technical Report. University of Texas at Austin.

[2] Abarja, R. A., & Wibowo, A. (2020). Movie Rating Prediction using ConvolutionalNeural Network based on Historical Values. *International Journal*, *8*(5).

[3] Dixit, P., Hussain, S., & Singh, G. (2020). Predicting the IMDB rating by using EDAand machine learning Algorithms.

[4] Sang-Ki Ko, Sang-Min Choi, Hae-Sung Eom , Jeong-Won Cha, Hyunchul Cho,Laehyum Kim, and Yo-Sub Han: A Smart Movie Recommendation System Content-based method uses item-to item similarity. If a user like B, we recommend A that is similar to B.

[5] Yibo Wang, Mingming Wang, and Wei Xu: A Sentiment-Enhanced Hybrid Recommender System for Movie Recommendation. As mentioned before, this paper uses collaborative filtering and content-based hybrid recommender systems.