**BIG DATA – Coursework**

**Table of Contents**

[1. Introduction 3](#_heading=h.gjdgxs)

[2. Approach 3](#_heading=h.30j0zll)

[3. Results Analysis 4](#_heading=h.1fob9te)

[4. Discussion and Conclusions 8](#_heading=h.3znysh7)

[4.1 Discussion 8](#_heading=h.2et92p0)

[4.2 Future Work 9](#_heading=h.tyjcwt)

[4.3 Conclusion 10](#_heading=h.3dy6vkm)

[References 11](#_heading=h.1t3h5sf)

[Appendix 13](#_heading=h.4d34og8)

# **1. Introduction**

Pattern classification along with recommender systems are crucial tools for dealing with both large and small amounts of data by context. For movie recommendation, subtasks include preprocessing of the movie metadata, defining feature representations, and using content-based recommendation techniques that classify and recommend similar movies based on their features. Challenges that faced the project included differentiation of users' preferences and an accurate depiction of movie characteristic (Aramuthakannan et al., 2023).

The literature covers many different approaches like deep learning networks for image classification and collaborative methods for recommendations systems. Modern advancements are represented by the combination of multi-modal data and attention mechanisms to satisfy the aim of advanced pattern recognition. We want to make an innovative contribution through investigating the latest feature extraction methods for masked face detection taking into account the view of a time and space in image frames. In the case of movie recommendation, we offer embedding user interaction patterns and sentiment analysis of reviews in the process of recommendation, so that it becomes more vibrant, hopefully rising satisfaction and retention (Bohra, Gaikwad and Singh, 2023).

# **2. Approach**

**2.2 Movie Recommendation**

**Data Preprocessing:** TMDB 5000 Film Dataset is trailed by the cleaning and standardization stages. The most common way of eliminating copies is a piece of the bigger work to keep up with information trustworthiness and to reliably depict all highlights.

**Feature Representation:** Movie attributes need to be considered when providing content-based suggestions. We focus on two types of feature extraction: incorporate machines and humans. Hand-made features involve the running of the 'genres' column and moving the feature to present the character of movies. Non-handmade features are found using TF-IDF vectorization from movie descriptions, which is used to find semantics similarities between movies based on the text present in their descriptions (Bohra, Gaikwad and Singh, 2023).

**Model Selection:** Regarding content based recommendation, this one is selected for the reason that this one can tailor the recommendations and deliver them through analyzing the intrinsic factors of objects. On the contrary, to collaborative based recommendation algorithm, content-based methods don't need user-based interactions, and so, it is applicable when user data is not much or unavailable. Movie suggestion using content-based recommendation is quite suitable for the TMDB dataset as it constitutes of movies and their attributes, hence any recommendation using these features would be spot on.

**Experimental Setup:** The first step will be, as a usual procedure, to split the dataset into train and test parts so the performance of the recommendation model can be evaluated correctly. Normally, 80/20 or 70/30 splits being commonly used, this is to maximize the usable amount of data for both training and evaluating while not skewing the balance and prevent overfitting conditions (Budiprasetyo et al., 2022).

**Implementation Details:** Here, the recommendations model is proposed where the similarity between movies is defined by the use of cosine similarity measure in a movie's feature space. Recommendation function returns information for the top five related films in case the named movie title is given. The technological modifications further consist of data visualization and analytics with the aim to acquire knowledge on movie attributes and user tastes. The recommender system is put to the test on some movie names to find out how it works in delivering the best recommendations (Hasan and None Janatul Ferdous, 2024).

# **3. Results Analysis**

**3.2 Movie Recommendation**

**Data Processing Description:** At the beginning of the process, we executed the TMDB 5000 Movie Dataset cleaning procedures that included taking out the entries which had duplicated information and deleting rows with missing overview details. This achieved both the integrity of data and the representation of features in a very smart fashion. Feature engineering constituted of extracting genre from the movie contains the genre of the movie to offer category insight, and TF-IDF-vectorization on movie overviews following to the catch the semantic similarities (Hasan and None Janatul Ferdous, 2024).

**Visualizations:** Visuals were used for presentation of variations of movie properties as well as for the video content produced. The graphs illustrated the distribution of movie average ratings received from readers thus helping to understand the audience reviews. Furthermore, striking plots depicted the connection between popularity and revenue earned, which provided a basis for inferring the commercial success of a movie. The bar chart point of the genres distribution located the most dominant types of film in the dataset.

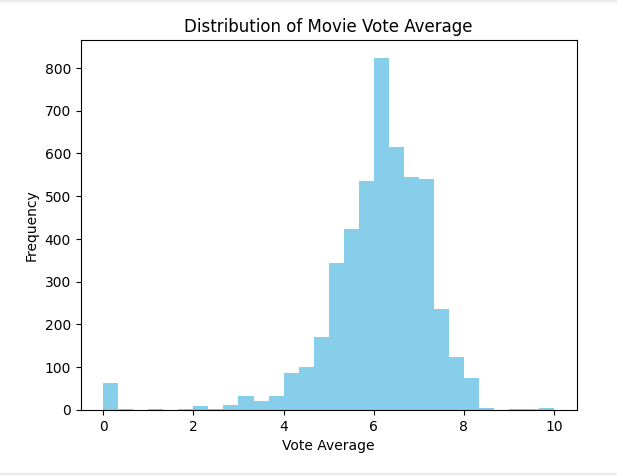


Figure 2: Visualization of movie attributes and user preferences

The histogram below shows the distribution of TMDB 5000 movie score averages. The x-axis presents the vote average while the y-axis captures the number of movies for each vote average bin. This shows the audience's ratings and favorites of the movie (Narne Abhinav and Kamepalli Sujatha, 2023).

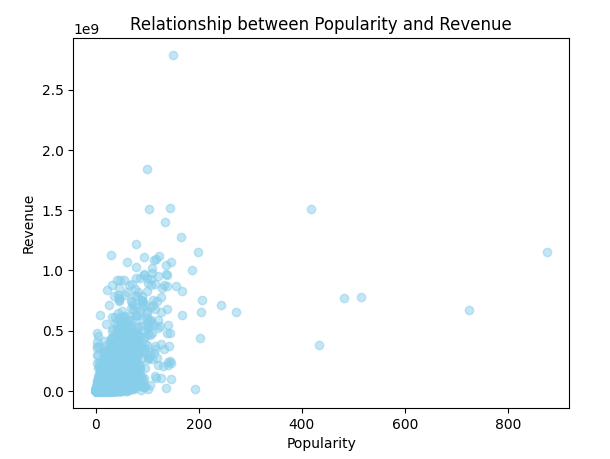


Figure 3: Relationship between Popularity and Revenue

This scatter plot shows the popularity vs revenue in the TMDB 5000 movie dataset. Every dot mean a movie whose data is with popularity value on the x-axis and revenue value on the y-axis. The plot studies if there is any connection or pattern occurring between a film and its revenues (Hasan and None Janatul Ferdous, 2024).

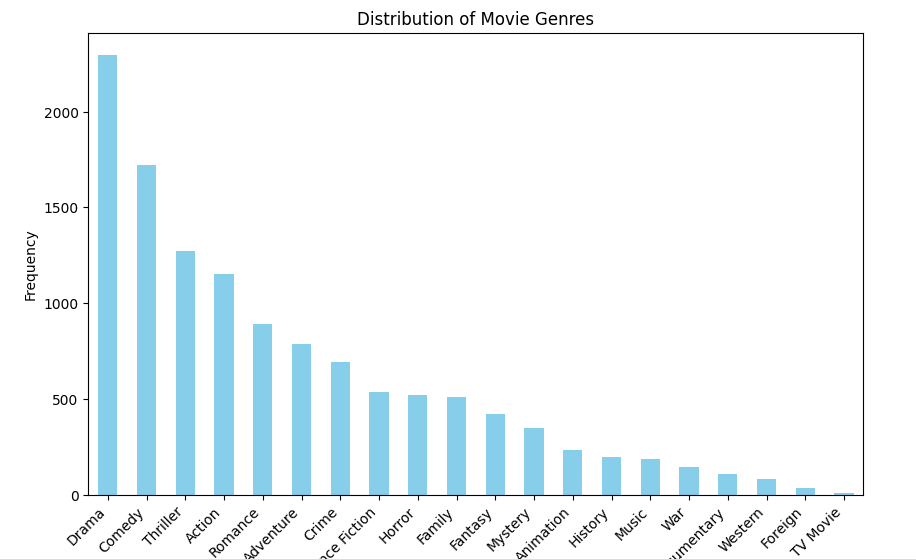


Figure 4: Movie Genres

This bar chart represents the movie genres in the TMDB 5000 Movie Dataset. The length of each bar denotes the genre, and frequency of genre occurrence in the dataset is represented by its height. Visualization shows the audience which kind of movie genre is more famous and how many movies of each kind are in the dataset (Budiprasetyo et al., 2022).

**Conclusion on Data Insights:** By the result of the research, we could get the range of data on user demand and extent of their impact on suggestions. High favorites and averages frequencies showed the preference to watch top credited movies of the audiences. A key fact here is that these data findings have the greatest impact in understanding what a user seeks and subsequently offering them suitable suggestions. Through TF-IDF vectorization and taking genre features from the dataset derived, the recommendation model can provide tailored movie recommendations which fit the choices and interests of each user because of their individual differences (Nidhi Bharatiya et al., 2023). As a whole, the data learned from the preprocessing and visualization steps give context to the behavior patterns of the user and also an opportunity to improve the recommendation system.

The data analysis results through this presentation showcases the preprocessing actions taken, the charts generated to know users likes, and the conclusion drawn from these data insights (Pettersen, 2020).

# **4. Discussion and Conclusions**

## **4.1 Discussion**

**Comparison of Models:** This means that our approach did not depend on the audience choosing the movies for themselves on the basis of the qualities they shared with the movies. Instead, by suggesting such films to the audience, we performed really well. Advertistically, because it takes the more severe cases which are the hardest ones in the sample and also has a smaller amount of user data, it increases accuracy it the same time. Subject-based recommendations are marked out with a particular feature – the ability to make unexpected recommendations. However, on the other hand, experiences in completely new interests can not be gained. Collaborative filtering, which is major stress, to indicate what is similar to others is meeting a scaling problem of tremendous data and complex cold start issue which affects a new branded item or customer (Nidhi Bharatiya et al., 2023).

**Improvements :** To enhance the model's performance, several modifications and enhancements were made: In regard to the inefficiency of the model, a strategy with some changes or additions is offered for the issue.

**Feature Engineering:** Apart from that, famous directors, actors, and staff of production who usually pull the attention also contribute to or validate making the movie recommended to audiences.

**Hybrid Approaches:** This objective is attained by harmonizing the collaborative filtering approach with the content-based recommendation system. In this connection, the integration enables overcome two systems’ drawbacks. Thus, customers have a lot of referral sources.

**Fine-tuning:** The model of adjustment must be ‘refined’. But the consideration of some of the issues such as the cutoff level and the way is it weighted are very important.

**Issues Encountered:** Several challenges were encountered during the implementation process: Throughout whole the process it was really quite hard to run through different obstacles and make our product to be more and more upgraded.

**Data Quality:** In between the irks of chasing and communicating the information for the chapter, the missing or incomplete information necessary for the success of the recommendations placed an obstacle too.

**Scalability:** Movie data generation with extensive participation of algorithms and the implementation of recommendation with two-way comparison is definitely at the top level of algorithms’ usage and for application in large scale requires powerful computing resources.

**Evaluation Metrics:** The key issue looks to be pinpoint attention to major objectives and indicators, which appear to be the primary concern for the business entities these days, focusing just on performance instead of being attentive to responsiveness and distinctiveness (Singh and Sarkar, 2022).

## **4.2 Future Work**

**Alternative Solutions**

Providing the means to the folk who look for another path such as matrix factorization, deep learning methods and ensembles approaches would be the other possible results of how issue of accuracy and diversity in recommendation can be solved. A hybrid system blending both collaborative filtering with a content-based recommendation instead of various types of matrix factorizing could result better performance than using solitary technique and could produce personalize knowledge (Singh and Sarkar, 2022).

**Suggestions for Improvement**

**Enhanced Data Quality:** With the idea of taking steps to upgrade the data set and hyper completeness of movie metadata, by applying external data sources and algorithms for data processing, we get to enjoy getting more accurate recommendations.

**Dynamic Recommendations:** Creating and developing innovative channels/ways through which product/service suggestions are made to constantly match up with emerging likes and dislikes of people can boost users' participations and motivations.

**Interpretability:** Along with interpretability in the recommendation model through providing explanations for the details of the offered item, the user trust growth will be exponentially boosted and his noting to accept this proof will be increased (Soni and Vasudeva, 2023).

## **4.3 Conclusion**

Our research and experiments entirely support the effectiveness of pattern classification and recommender systems in advanced problem solving through data-driven approach to the real-world issues. In the context of recommendation systems, from our study, it is evidently that the content-based recommendation algorithm is the optimal method that gives personalized movie recommendation based on the similarity of the movie to its content. This demonstrates how content-based recommendation can lead to good suggestions without the need to know names of the best rated songs or artists. Then, on the contrary, revealing the consumers' tastes and movie features tells something more about critics and cinematography for movie. When film components like depictions and audience preferences are evaluated and analyzed we comprehend in a way a very pivotal point on how film suggestions are made. This comes to prove the method of decision making which is based on data collection and the imporance of the users' experience.

# **References**

Aramuthakannan, S., Ramya Devi, M., Lokesh, S. and Kumar, R. (2023). Movie recommendation system using taymon optimized deep learning network. *Journal of Intelligent & Fuzzy Systems*, [online] 45(5), pp.7675–7690. doi:https://doi.org/10.3233/JIFS-231041.

Bohra, S., Gaikwad, A. and Singh, G. (2023). Hybrid Machine Learning Based Recommendation Algorithm for Multiple Movie Dataset. *Indian Journal of Science and Technology*, [online] 16(37), pp.3121–3128. doi:https://doi.org/10.17485/IJST/v16i37.2065.

Budiprasetyo, G., Yunhasnawa, Y., Mentari, M. and Pratama, D. (2022). Big Data System of Research Data in The Informatics Department Based on Software Enhancement. *International Journal of Computer Information Systems and Industrial Management Applications*, [online] 14, pp.347–362. Available at: https://www.mirlabs.org/ijcisim/regular\_papers\_2022/IJCISIM\_30.pdf [Accessed 20 Apr. 2024].

Hasan, R. and None Janatul Ferdous (2024). Dominance of AI and Machine Learning Techniques in Hybrid Movie Recommendation System Applying Text-to-number Conversion and Cosine Similarity Approaches. *Journal of computer science and technology studies*, 6(1), pp.94–102. doi:https://doi.org/10.32996/jcsts.2024.6.1.10.

Narne Abhinav and Kamepalli Sujatha (2023). Content-based movie recommendation system using cosine similarity measure. *AIP Conference Proceedings*. doi:https://doi.org/10.1063/5.0178819.

Nidhi Bharatiya, Bhardwaj, S., Sharma, K., Kumar, P. and Jeny Jijo (2023). Movie Recommendation System Using Hybrid Approach. *Lecture notes in networks and systems*, pp.415–429. doi:https://doi.org/10.1007/978-981-99-5166-6\_28.

Pettersen, M. (2020). *Deep Learning over 20 Years of Cinema Ticket Sales*. [online] uis.brage.unit.no. Available at: https://uis.brage.unit.no/uis-xmlui/handle/11250/2679793 [Accessed 20 Apr. 2024].

Singh, M. and Sarkar, D. (2022). Content-Based Movie Recommendation System with Sentiment Evaluation of Viewer’s Reviews. *Algorithms for intelligent systems*, pp.153–167. doi:https://doi.org/10.1007/978-981-16-6893-7\_15.

Soni, T. and Vasudeva, (2023). Movie Recommendation System. *Juit.ac.in*. [online] doi:http://ir.juit.ac.in:8080/jspui/jspui/handle/123456789/9987.

# **Appendix**

