Find Films that Match Your Vibe

MovieAura

Group 17

Nikola Raicevic/ Rishabh Kumar/ Wendi Tan/ Yifan Peng/ Yi Hsuan Wen

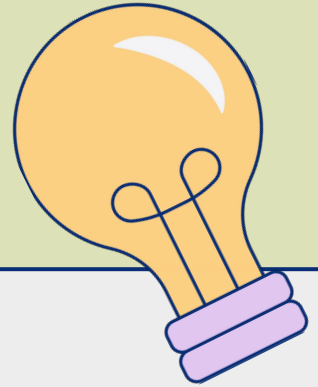
Outline

1. Motivation and Objectives
2. Database
3. Methodology
4. System Architecture
    a. Submodule 1
    b. Submodule 2
    c. Submodule 3
    d. Submodule 4
5. Result & Strengths of the model
6. Conclusion

Outline

1. Motivation and Objectives
2. Database
3. Methodology
4. System Architecture
   a. Submodule 1
   b. Submodule 2
   c. Submodule 3
   d. Submodule 4
5. Result & Strengths of the model
6. Conclusion

Motivation and Objective



Motivation

Streaming platforms' vast libraries make it hard for users to find movies that match their tastes

Approach

Use collaborative and content-based filtering, integrating user history with movie metadata for recommendations.

Impact

Simplify content discovery, reduce search time, and boost user engagement and retention.

Outline

1. Motivation and Objectives
2. Database
3. Methodology
4. System Architecture
   a. Submodule 1
   b. Submodule 2
   c. Submodule 3
   d. Submodule 4
5. Result & Strengths of the model
6. Conclusion

Movie Dataset

Dataset 1

Full TMDB Movies Dataset 2024

- Time frame: ~ 2024

Dataset 2

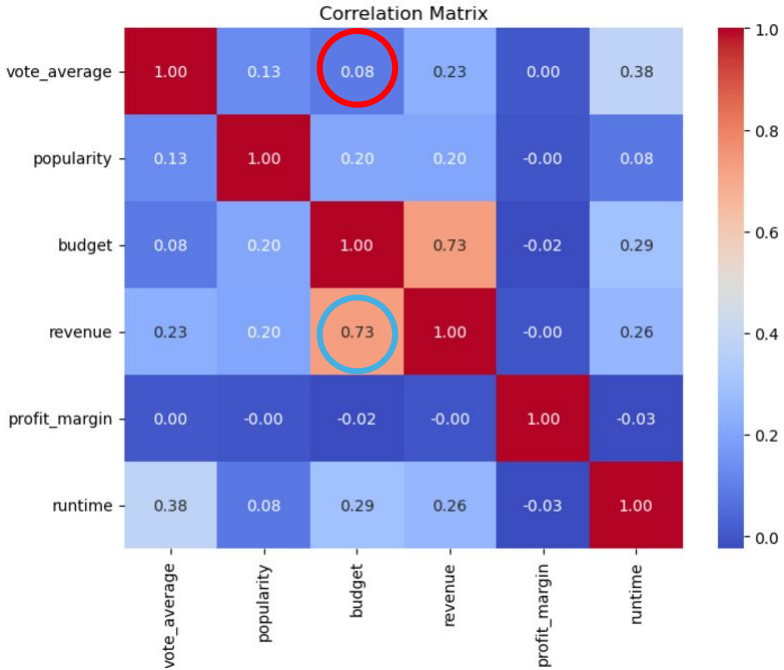
16000+ Movies (Metacritic)

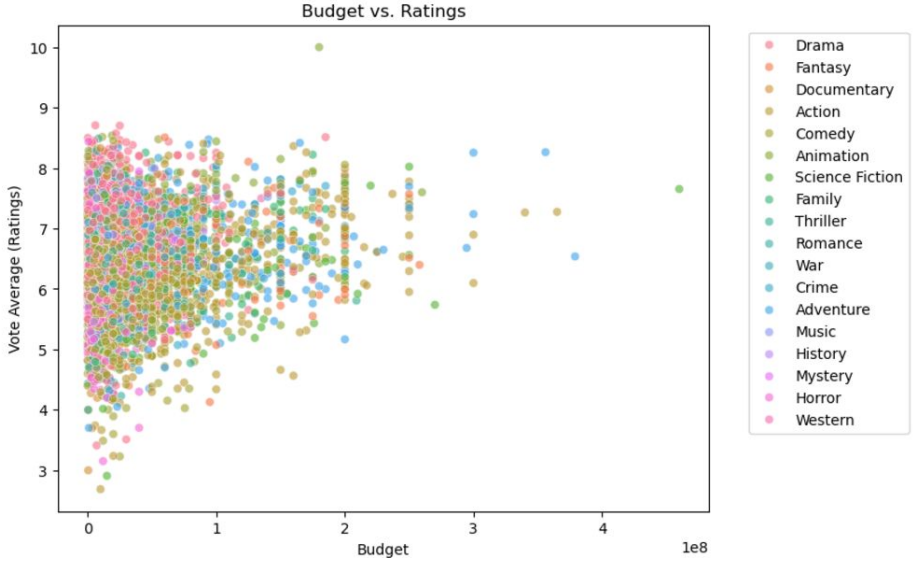
- Time frame:1910-2024

Our Dataset

- Entries: 16,000+ movies
- Features: 24+ columns, including titles, release years, genres, revenue, popularity, runtime, ratings, and posters
- Data Processing: Integrated datasets for combined insights; content-based features analyzed for trend and prediction modeling

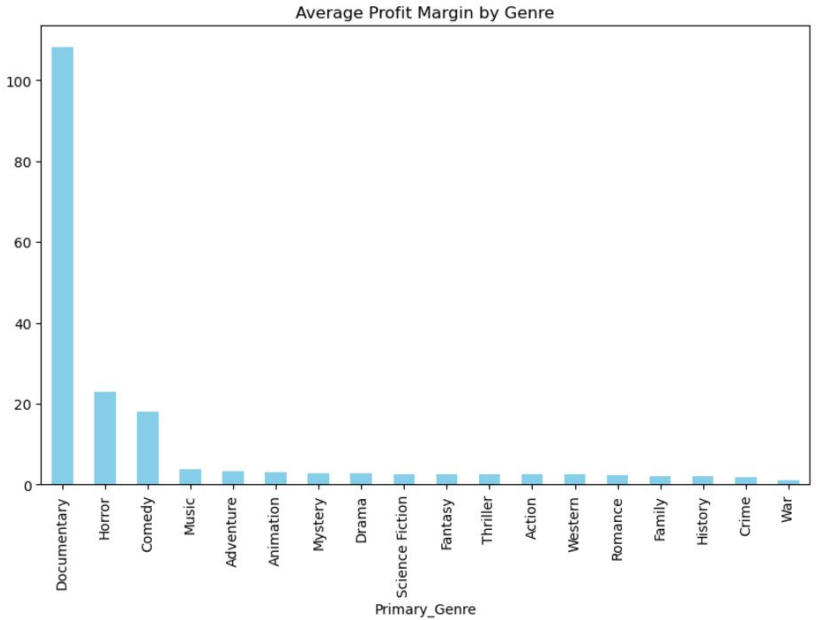
Higher Budgets Don't Always Mean Better Ratings, But They Can Drive Revenue

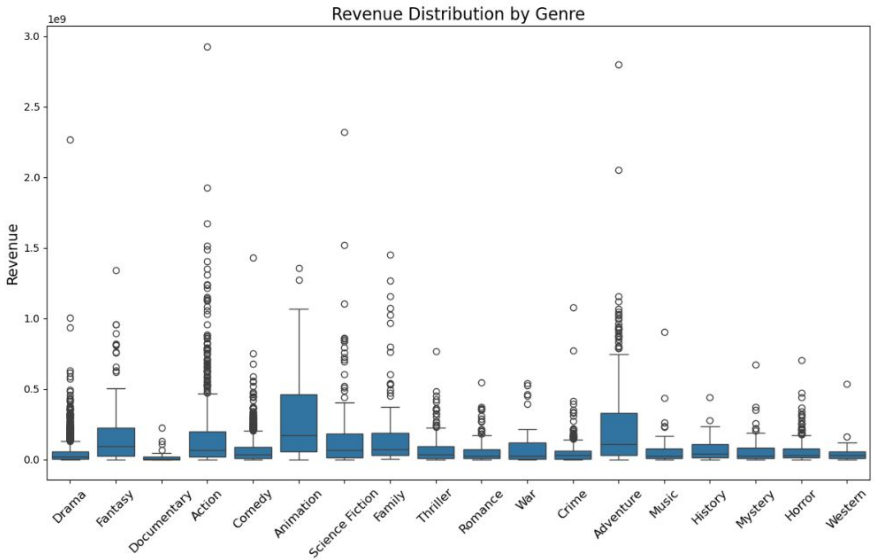




- Higher budget does not guarantee better ratings, smaller films with strong storytelling or niche appeal succeed critically.
- Budget and revenue have +ve correlation, higher-budget movies tend to generate more revenue.

Revenue and Profit vs Genre





- Documentary movies have highest Profit Margin.
- Animation Genre have higher revenue in general, Action Genre have more outliers towards higher revenue side

Outline

1. Motivation and Objectives
2. Database
3. Methodology
4. System Architecture
   a. Submodule 1
   b. Submodule 2
   c. Submodule 3
   d. Submodule 4
5. Result & Strengths of the model
6. Conclusion

Methodology

01. Data Collection and Preprocessing:
    a. Data from various sources was merged by matching movie titles. Features like genres, keywords, and release years were extracted and standardized for submodule input.
02. Weighting and Aggregation:

    Weights are assigned to submodules based on their relative importance and output scores

    a. Submodule_01 (Image-Based): w1 = 50
    b. Submodule_02 (Content-Based): w2 = 10
    c. Submodule_03 (Graph-Based): w3 = 1
    d. Submodule_04 (Transformer): w4 = 100
03. Final scores are computed as a weighted sum of submodule outputs. The top five recommendations are displayed based on the final aggregated scores:

    Final Score= w1×Sub_01 + w2×Sub_02 + w3×Sub_03 + w4×Sub_04

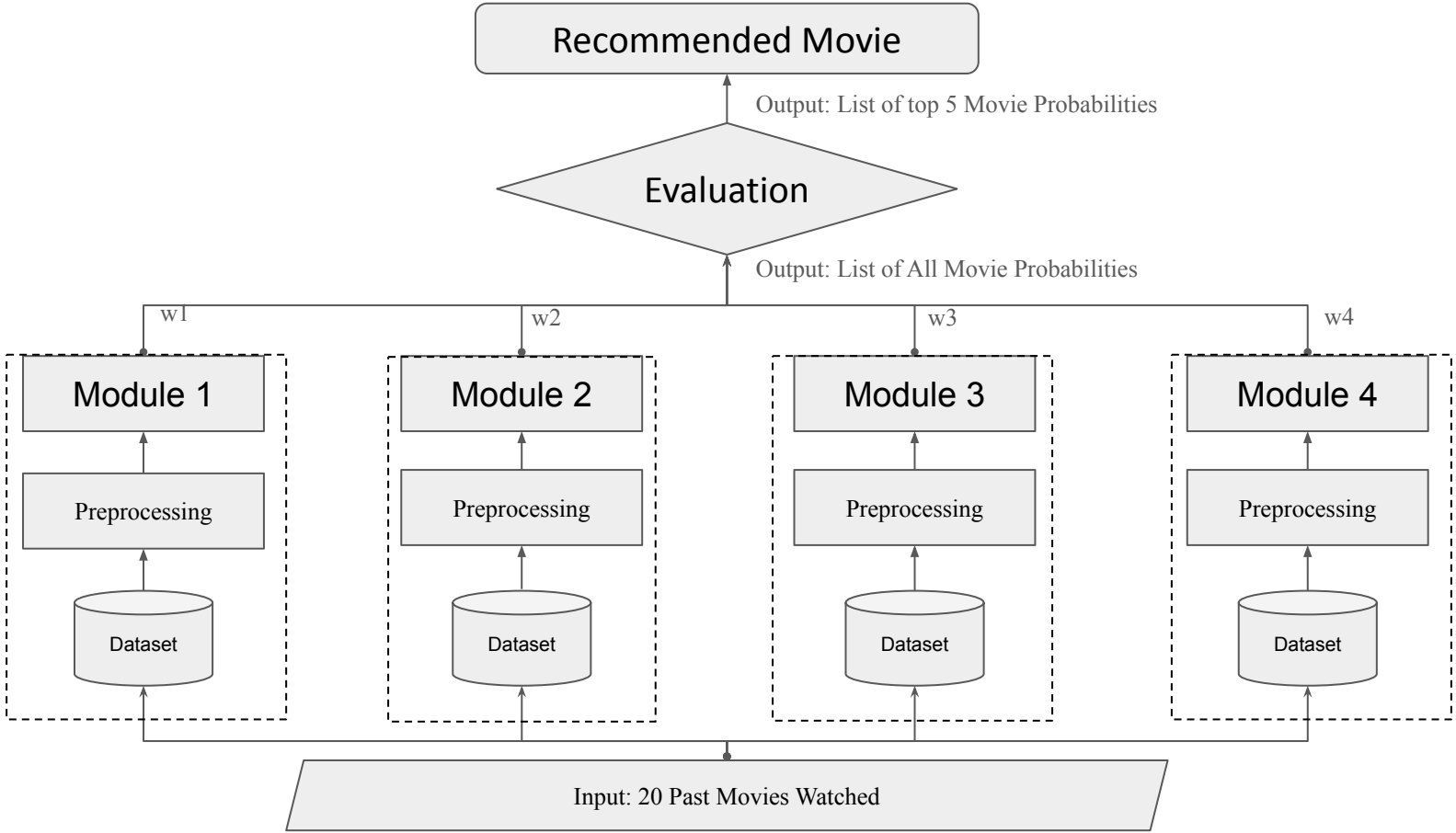
Methods

1) Image Classification
    a) CNN (pre-trained ResNet-50) combining with content-based filtering
2) Content Based Filtering:
    a) Neural Collaborative Filtering (NCF)
    b) Multilayer Perceptron (MLP)
3) Collaborative Filtering:
    a) Similarity-Based Models
    b) Singular Value Decomposition (SVD)
    c) Alternating Least Squares (capture user-movie interactions by breaking down the user-movie matrix into lower-dimensional matrices)
    d) Latent Factor Models
4) LLms
5) Sequence Based Models
    a) Recurrent Neural Networks
    b) Long Short Term Memory
    c) Transformers
6) Graph Based Techniques
    a) Graph Neural Network
    b) Graph Based Collaborative Filtering
7) Reinforcement Learning
    a) Deep Q Learning
    b) Policy gradient methods

Outline

1. Motivation and Objectives
2. Database
3. Methodology
4. System Architecture
    a. Submodule 1
    b. Submodule 2
    c. Submodule 3
    d. Submodule 4
5. Result & Strengths of the model
6. Conclusion

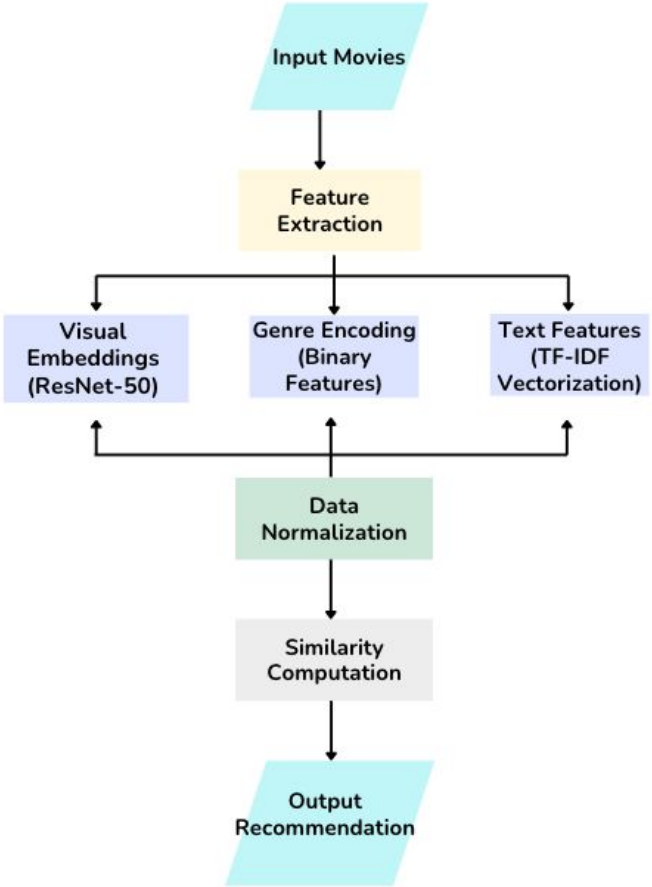
System Architecture



Outline

1. Motivation and Objectives
2. Database
3. Methodology
4. System Architecture
   a. Submodule 1
   b. Submodule 2
   c. Submodule 3
   d. Submodule 4
5. Result & Strengths of the model
6. Conclusion

Submodule 01: Visual & Content-Based Filtering

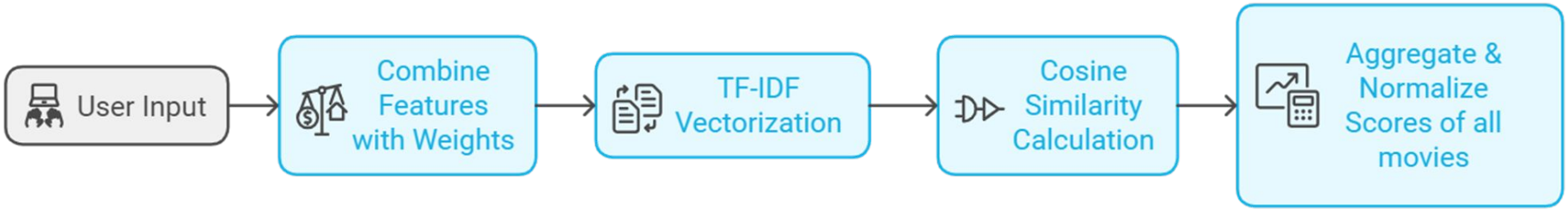


- **Feature Extraction:**
  Combined visual embeddings (using pre-trained ResNet-50 to extract movie posters features), genre features (binary encoding), and text features (TF-IDF on descriptions).
- **Data Normalization:**
  Scaled all features (visual, genre, and text) to a uniform range for integration.
- **Similarity Computation:**
  Generated a similarity matrix using cosine similarity on combined features.
- **Hybrid Recommendation System:**
  Matched input movies with the most similar titles based on combined visual, textual, and genre similarity.

Outline

1. Motivation and Objectives
2. Database
3. Methodology
4. System Architecture
   a. Submodule 1
   b. Submodule 2
   c. Submodule 3
   d. Submodule 4
5. Result & Strengths of the model
6. Conclusion

Submodule 02: Content Based Filtering



- **Feature Combination and Weighting**: Combine key features like genres (×2), keywords (×3), and descriptions into a weighted text representation for each movie.

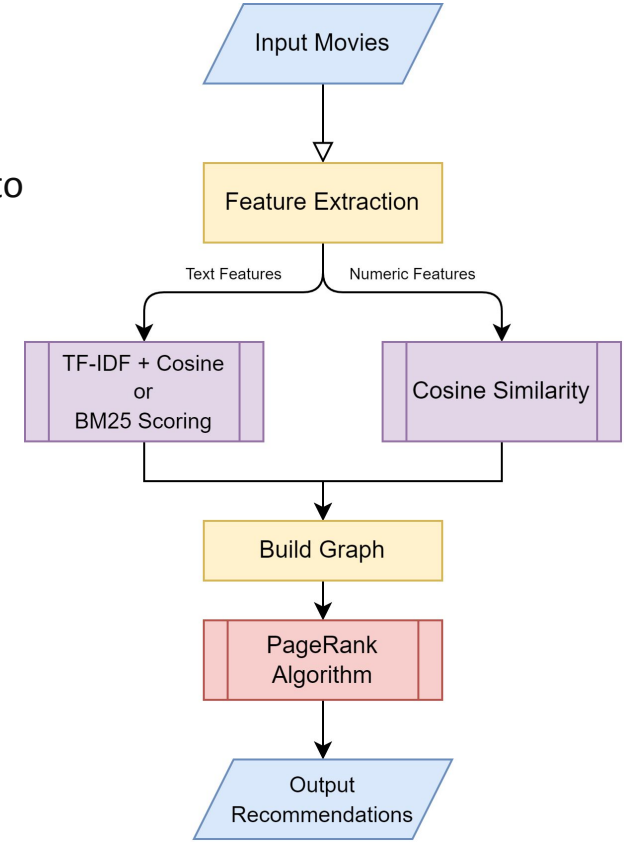
- **Similarity-Based Recommendation:** Use TF-IDF vectorization and cosine similarity to calculate and rank movies based on their relevance to user-selected titles.

Outline

1. Motivation and Objectives
2. Database
3. Methodology
4. System Architecture
   a. Submodule 1
   b. Submodule 2
   c. Submodule 3
   d. Submodule 4
5. Result & Strengths of the model
6. Conclusion

Submodule 03: Graph Based Filtering

- **Feature Extraction**: Incorporates text features (e.g., company, language) and numeric features (e.g., release
- **Text Mining**: Utilizes **TF-IDF vectorization** and **BM25 scoring** to analyze textual attributes.
- **Numeric Processing**: Normalizes numeric attributes using StandardScaler and computes similarity with **cosine similarity**.
- **Graph-Based Recommendation**:
  - Represents movies as graph nodes with edge weights based on combined text and numeric similarity.
  - Refines rankings iteratively via **PageRank (PR) Algorithm**, leveraging connectivity for both local and global relationships.



Outline

1. Motivation and Objectives
2. Database
3. Methodology
4. System Architecture
   a. Submodule 1
   b. Submodule 2
   c. Submodule 3
   d. Submodule 4
5. Result & Strengths of the model
6. Conclusion

Submodule 04: Transformer

- **Model Details:**

  The model used is SASRec/BERT4Rec Transformer-based architecture designed specifically for recommendation tasks. The model uses self-attention mechanism to model user sequence and can handle both sequential dependencies (past movies) and incorporate information or context based embeddings.

- **Features Extraction:**

  The features extracted includes the list of movies with their specific numerical embeddings including Adult movie, Genres, Production Country, Keywords, Description, Rating, Popularity, and Release Year.

  The Adult category was binary either true/false; Genres and Production Country was one-hot encoded; Rating, Popularity, and Release Year was normalized so that the features are in the range of 0 to 1; Keywords and Description features were extracted using pre-trained language model BERT to create a dense vector representation.

- **Input Features:**

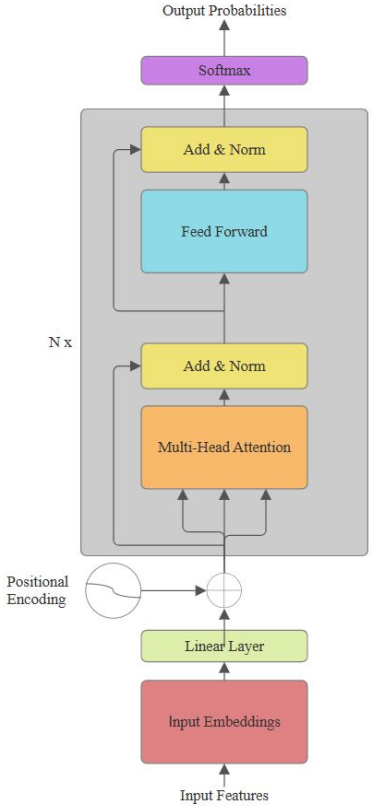
  Generated a sequence of a movie from the list paired with random SEQ_LEN - 1 movies (19 movies)

- **Transformer Based Recommendation:**

  The model uses sequence of 20 movies with 128 embeddings after linear layer which handles 1707 input features. The model itself has 2 layers and 4 self-attention heads used to find the cross-similarity across the movies.

- **Output Probabilities:**

  The transformer itself outputs logits which are converted into output probabilities once passed through the softmax layer.



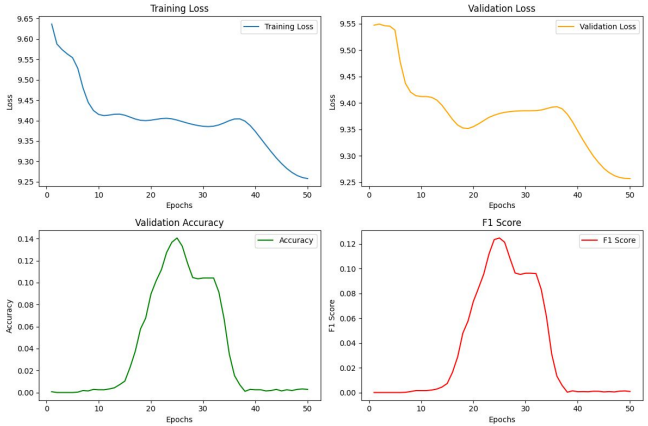
Submodule 04: Transformer Training/Evaluation

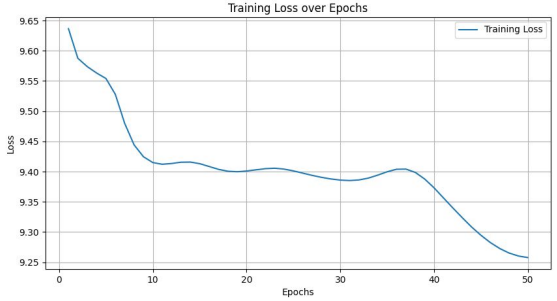
- **Training:**

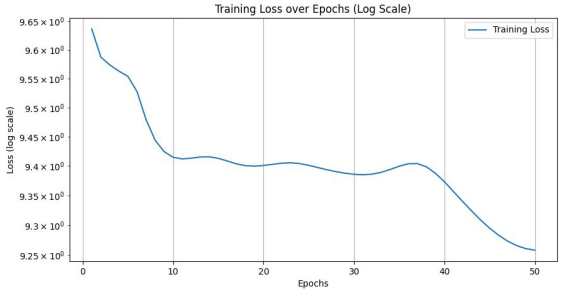
  The model is trained using the batch size of 32 on the dataset containing 14010 different movies for 50 epochs. The model of 30 epoch training was used as it was found to be the optimal model.

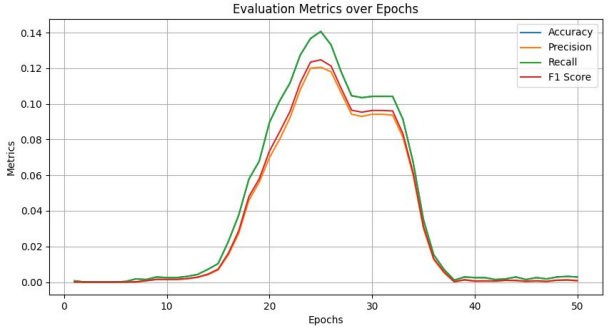
- **Evaluation:**

  The model was evaluated based on accuracy, precision, recall, and F1 score. The accuracy measures ratio of correctly predicted instances to the total number of instances. The precision measures ratio of true positive predictions to the total positive predictions made by the model. The recall measures the ratio of true positive predictions to all actual positives. The F1 measures the the harmonic mean of precision and recall.









Outline

1. Motivation and Objectives
2. Database
3. Methodology
4. System Architecture
   a. Submodule 1
   b. Submodule 2
   c. Submodule 3
   d. Submodule 4
5. Result & Strengths of the model
6. Conclusion

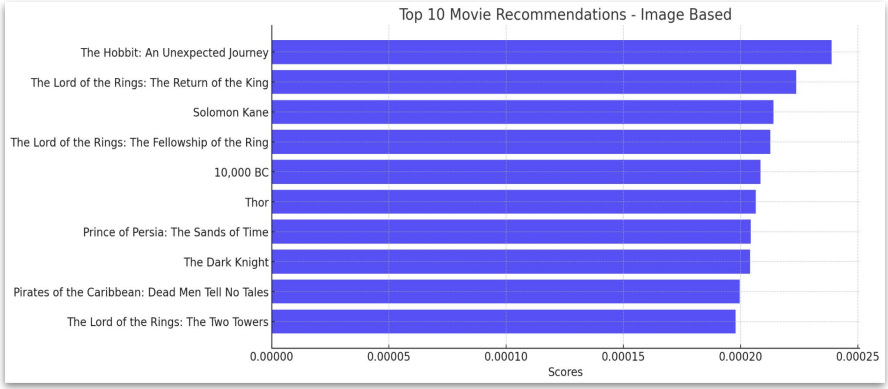
Results: Top 10 Movie Recommendations based on each submodule

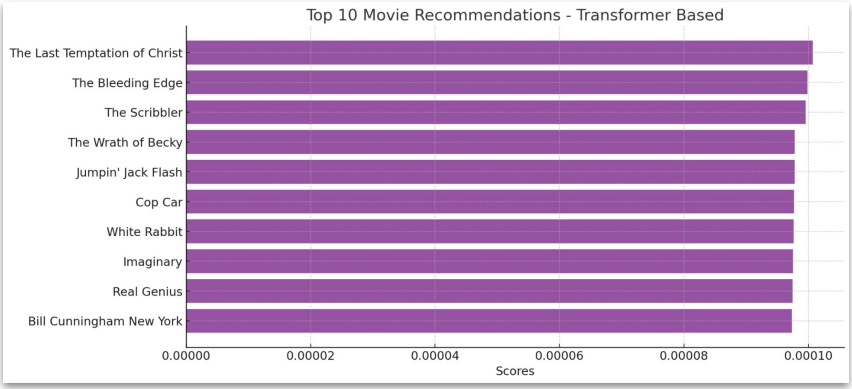
Input:

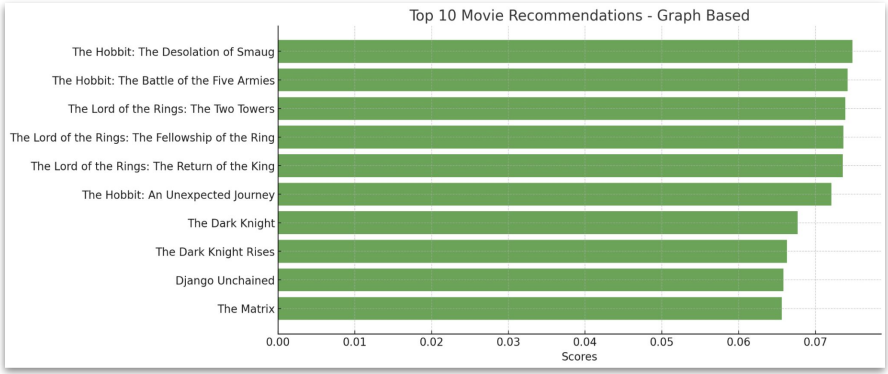
Movies movie_sequence = ['The Lord of the Rings: The Return of the King', 'The Lord of the Rings: The Fellowship of the Ring','The Lord of the Rings: The Two Towers','The Matrix', 'The Dark Knight Rises', 'Interstellar','Django Unchained', 'The Godfather', 'The Shawshank Redemption', 'The Dark Knight','The Hobbit: An Unexpected Journey', 'The Hobbit: The Desolation of Smaug', 'The Hobbit: The Battle of the Five Armies']

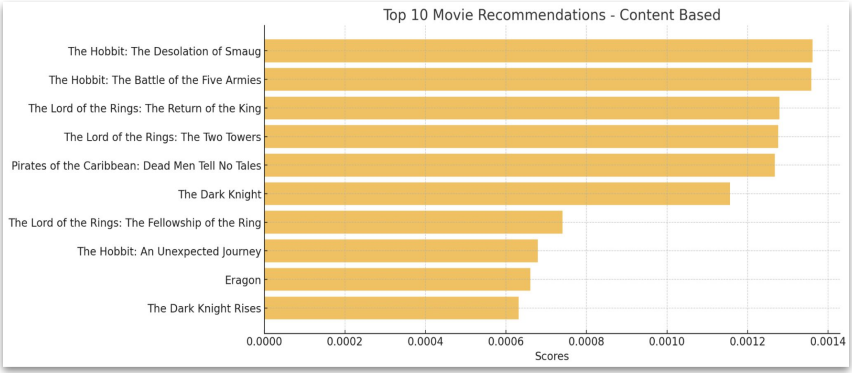
Graph Based		Image Based		Content Based		Transformer	
The Hobbit: An Unexpected Journey	0.00024	The Hobbit: The Battle of the Five Armies	0.00136	The Hobbit: The Desolation of Smaug	0.07481	The Last Temptation of Christ	1.0070e-04
The Lord of the Rings:The Return of the King	0.00022	The Lord of the Rings: The Return of the Kin	0.00135	The Hobbit: The Battle of the Five Armies	0.07420	The Bleeding Edge	9.9866e-05
Solomon Kane	0.00021	The Lord of the Rings: The Two Towers	0.00128	The Lord of the Rings: The Two Towers	0.07389	The Scribbler	9.9549e-05
The Lord of the Rings: The Fellowship of the Ring	0.00021	The Hobbit: The Battle of the Five Armies	0.00127	The Lord of the Rings: The Fellowship of the Ring	0.07362	The Wrath of Becky	9.7789e-05
10,000 BC	0.00021	The Lord of the Rings: The Return of the King	0.00126	The Lord of the Rings: The Return of the King	0.07357	Jumpin' Jack Flash	9.7772e-05
Thor	0.00021	The Hobbit: An Unexpected Journey	0.00115	The Hobbit: An Unexpected Journey	0.07207	Cop Car	9.7654e-05
Prince of Persia: The Sands of Time	0.00020	Eragon	0.00074	The Dark Knight	0.06768	White Rabbit	9.7594e-05
The Dark Knight	0.00020	The Dark Knight Rise	0.00067	The Dark Knight Rises	0.06629	Imaginary	9.7536e-05
Pirates of the Caribbean: Dead Men Tell No Tales	0.00020	Warcraft	0.00066	Django Unchained	0.06583	Real Genius	9.7429e-05
The Lord of the Rings: The Two Towers	0.00019	The Dark Knight	0.00063	The Matrix	0.06561	Bill Cunningham New York	9.7352e-05

Top 10 Movie Recommendations based on each submodule:

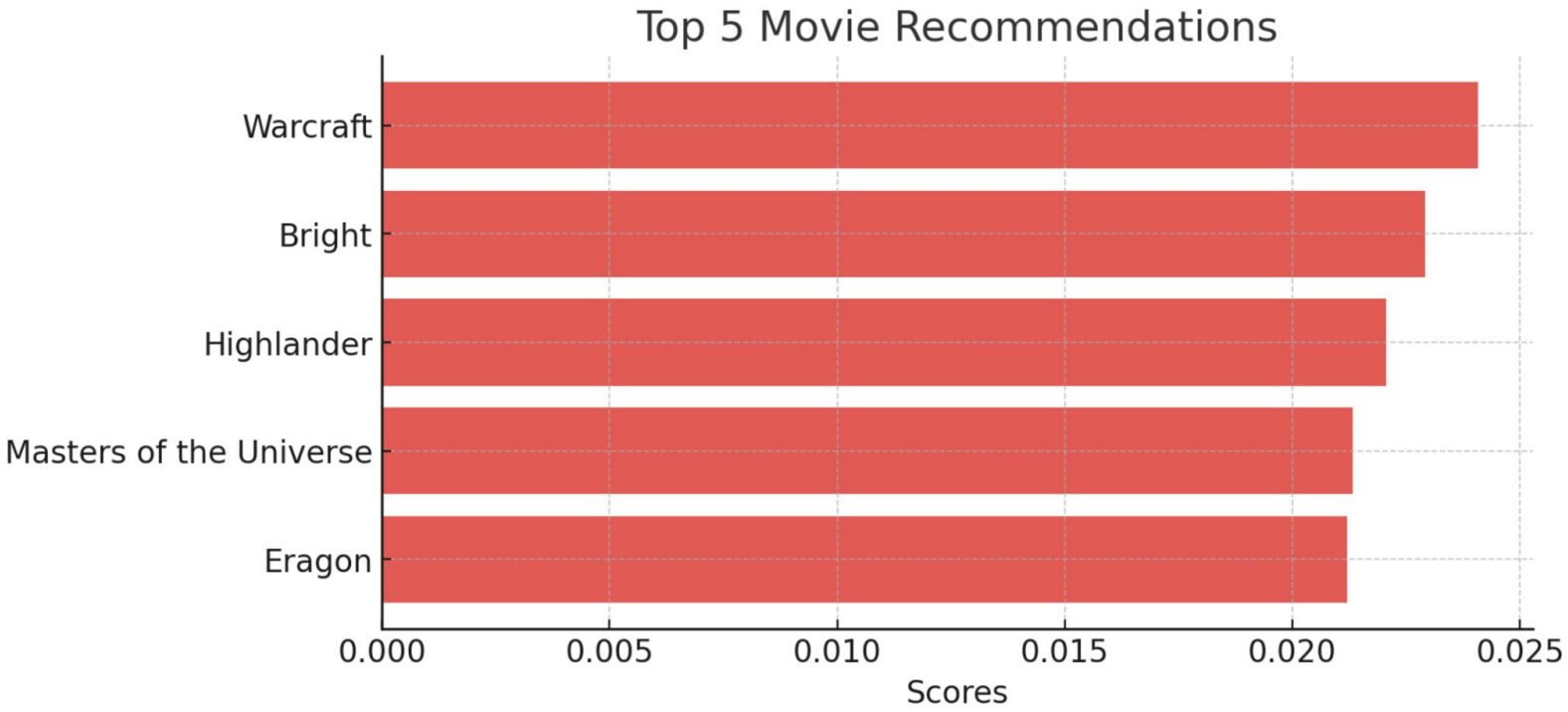




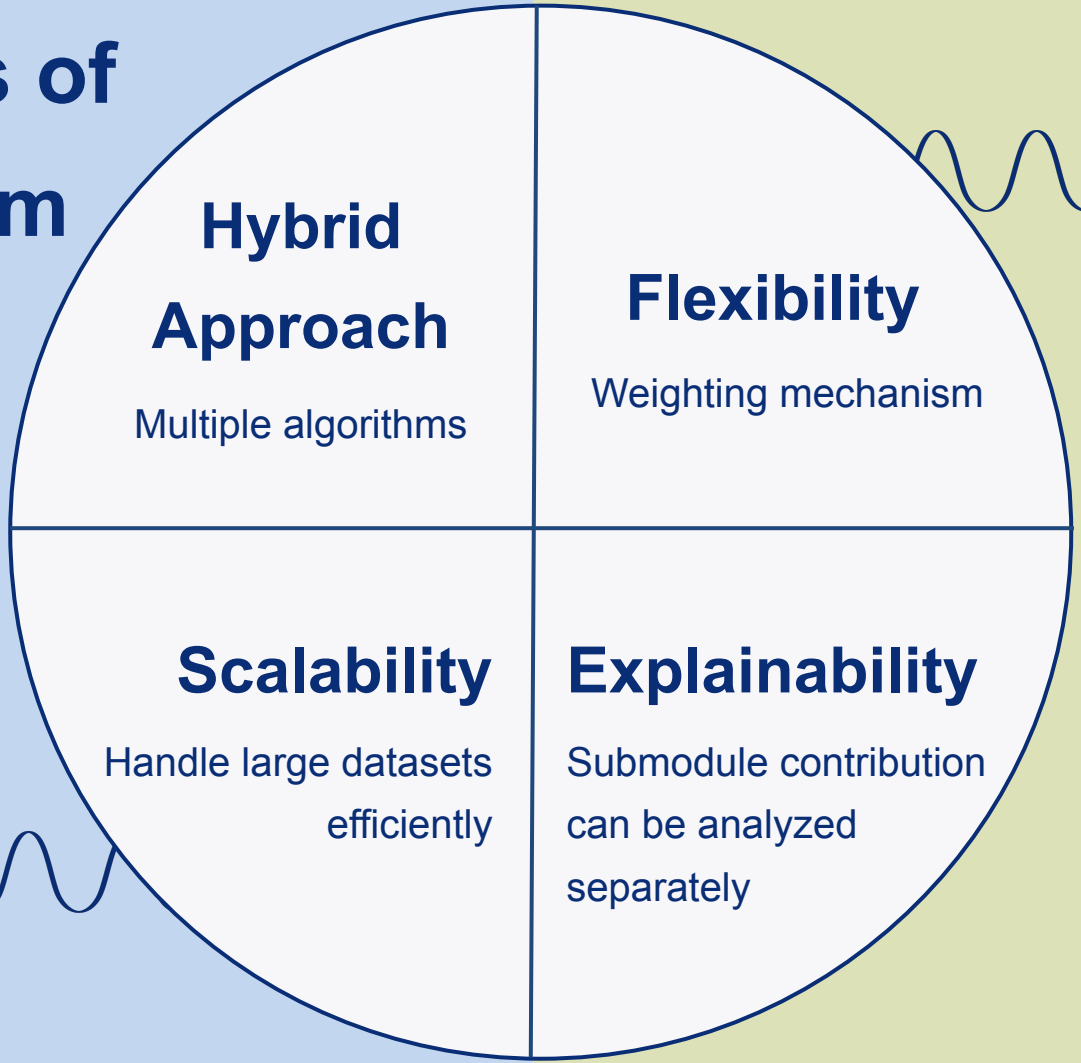




Top 5 Movie Recommendations:



Strengths of our system



Outline

1. Motivation and Objectives
2. Database
3. Methodology
4. System Architecture
   a. Submodule 1
   b. Submodule 2
   c. Submodule 3
   d. Submodule 4
5. Result & Strengths of the model
6. Conclusion

Conclusion

- **Holistic Recommendation Approach**: Combines content-based, image-based, graph-based, and transformer-based techniques for robust movie recommendations.
- **Data-Driven Insights**: Visualizations using Seaborn and Matplotlib provide a deeper understanding of trends and user preferences.
- **Scalability & Flexibility**: Modular architecture allows easy integration of additional algorithms and features.
- **Real-World Applicability**: Addresses the challenge of personalized content discovery in large datasets effectively.
- **State-of-the-Art Techniques**: Leverages advanced models like ResNet-50 and transformers for innovative solutions.
- **User-Focused Design**: Prioritizes relevance and usability to enhance the user experience.

Thank you!

Any questions?