MovieMate AI Movie Recommender

Souri Satya Saketh Ravi UID:121292803
Bodla Krishna Vamshi UID:121322456
Swapnita Sahu UID:121292223

Abstract

Recommender systems are essential for improving user experience on modern streaming platforms, yet generating meaningful suggestions remains challenging when user history is limited. To address this, we present MovieMate, a lightweight and interpretable movie recommendation system based on content-based filtering. Using the IMDB Top-1000 Movies dataset, we preprocess textual and numeric metadata including genres, directors, casts, runtime, and ratings to construct TF-IDF feature vectors that capture semantic similarity among films. We then compute cosine similarity to rank recommendations without requiring user profiles or interactions. Exploratory data analysis reveals critical trends in movie characteristics, informing our feature selection and validating our focus on descriptive attributes. Our results demonstrate that MovieMate can rapidly produce accurate and explainable recommendations while remaining scalable to larger catalogs. Although the absence of visual/audio embeddings limits personalization and stylistic nuance, our system establishes a strong baseline for future hybrid models that incorporate user preference learning and multimodal representations

1 Introduction

The rapid growth of online streaming platforms has made it increasingly challenging for users to efficiently discover content aligned with their viewing preferences. Traditional search and browsing interfaces often struggle to surface relevant items from large scale catalogs, leading to decision fatigue and reduced engagement. Recommender systems address this challenge by enabling personalized content discovery.

While collaborative filtering methods depend on historical user interactions, such data are often sparse, unavailable for new users, or constrained by privacy considerations. To address these limitations, we investigate a content-based recommendation approach that requires no behavioral data. We propose *MovieMate*, a movie recommender that models film similarity using descriptive metadata from the IMDB Top1000 Movies dataset, including genre, cast, director, runtime, and user ratings. These attributes are preprocessed to ensure consistency and reduce noise.

Movie representations are constructed using TF-IDF vectorization over textual features, with cosine similarity used to rank recommendations. This design offers key advantages, including scalability, interpretability, and robustness to limited user information. Exploratory data analysis further informs system design by revealing dataset characteristics such as genre dominance and the moderate influence of popularity metrics.

Although the approach is limited by its reliance on textual metadata and cannot fully capture stylistic or multimodal aspects, MovieMate provides a strong, extensible foundation. Future work may incorporate hybrid models, embedding-based retrieval, or personalization mechanisms to support richer recommendation scenarios.

2 Related works

Recommender systems have been extensively studied across domains such as e-commerce, streaming platforms, and social media, with two primary paradigms dominating the literature: content-based filtering and collaborative filtering. Collaborative methods rely on user–item interaction data to learn preference patterns but suffer from cold-start and sparsity limitations. In contrast, content-based approaches compute similarity directly between items

using descriptive features such as text metadata, enabling interpretable and user history independent recommendations. This makes content-based modeling a natural fit for scenarios where user profiles are limited or new users are continuously introduced. MovieMate adopts this approach by representing films with TF-IDF vectors derived from textual metadata including genres, directors, and cast lists.

Prior work in movie recommendation has explored rich multimodal inputs such as plot embeddings, trailer audio visual features, and user sentiment to improve personalization. However, these methods typically require large training datasets and significant inference resources. Lightweight similarity-based models remain highly valuable when efficiency, transparency, and ease of deployment are prioritized. MovieMate aligns with this design philosophy and builds upon the foundation of established similarity-based recommendation frameworks, showing that meaningful results can be achieved even when restricting features to structured metadata and text analysis.

3 Dataset and pre processing

Our system is developed using the publicly available IMDB Top-1000 Movies dataset, which provides curated information for critically acclaimed and widely recognized films. The dataset includes structured metadata such as title, genre, director, main cast, runtime, release year, IMDB rating, number of votes, and gross revenue. These features collectively describe narrative attributes, stylistic influences, and popularity trends that can be leveraged for similarity-based recommendations. Before modeling, we identified several inconsistencies arising from formatting artifacts and missing data. For instance, commas, currency symbols, and alphabetic suffixes embedded within numeric fields (e.g., “min”, “$”) disrupted type integrity and were removed to ensure proper conversion to numerical formats.

We applied a targeted data-cleaning strategy, including coercion of invalid numeric values to NaN, followed by median or mode imputation for missing entries depending on attribute type (e.g., runtime vs. genre). Runtime and year were standardized to numeric form, and categorical textual fields (e.g., directors, actors) were preserved for later feature engineering. These preprocessing steps enhanced dataset reliability, removed noise, and ensured uniform representation across attributes, ultimately enabling more robust and interpretable recommendation outcomes.

4 Exploratory Data Analysis

To understand the dataset’s representational dynamics, we conducted exploratory data analysis (EDA) focusing on distributional patterns and feature correlations. Rating histograms revealed that most films cluster tightly between 7.7 and 8.2, suggesting consistently high quality and limiting the discriminative utility of ratings alone. Genre frequency analysis demonstrated a strong dominance of drama, revealing a potential genre-driven bias in similarity recommendations and underscoring the need for richer multi-attribute modeling. We further evaluated temporal trends, showing a significant increase in film production post 1990, correlated with industry globalization and digital transformation.

Scatter and heatmap-based correlation analyses indicated only modest relationships between key variables runtime, rating, and release year while popularity driven metrics such as votes and gross revenue were moderately correlated with each other but not strongly with quality indicators. This confirmed that textual metadata (e.g., genre, director, cast) provides more meaningful semantic structure for similarity measurement than purely numeric descriptors. The insights derived from EDA guided feature prioritization and validated the selection of a content based modeling strategy for effective recommendation performance.

5 Method Implementation

MovieMate is implemented as a content-based, item–item recommendation system that operates entirely on movie metadata and does not require user interaction data. Each movie is represented using textual descriptors extracted from the dataset, including genre labels, director names, cast members, and plot summaries. These fields are concatenated into a single document per movie and encoded using TF-IDF vectorization, producing high-dimensional sparse representations that emphasize distinctive lexical cues while down-weighting ubiquitous terms.

Movie similarity is computed using cosine similarity, which measures angular proximity in the TF-IDF vector space and enables efficient ranking of candidate recommendations. This formula-

tion allows MovieMate to capture semantic overlap arising from shared themes, creative contributors, and narrative elements while remaining fully interpretable. Importantly, the modular design of the feature pipeline allows individual metadata fields to be added or removed without changing the underlying similarity computation, enabling controlled ablation and comparison experiments.

This approach ensures fast inference, transparency in recommendation behavior, and robustness to cold-start scenarios commonly encountered in collaborative filtering systems. While intentionally lightweight, the architecture provides a strong and reproducible foundation upon which more expressive hybrid or multimodal recommender models can be built.

6 Experimental Design

To evaluate the effect of metadata richness on recommendation quality, we design a controlled comparison between a minimal baseline model and an enhanced combined-metadata model. This comparison satisfies the requirement for a clear baseline and allows us to isolate the contribution of additional linguistic features. Both models use identical preprocessing, TF-IDF configurations, and cosine similarity computations; they differ only in the metadata fields included during vectorization.

Because the task is unsupervised and no ground-truth relevance labels are available, evaluation focuses on comparative similarity behavior, qualitative coherence of recommendations, and summary statistics rather than predictive accuracy.

6.1 Baseline Model: Genre-Only TF-IDF

The baseline model represents each movie using only its genre labels. Genres are tokenized and encoded using TF-IDF, and cosine similarity is used to generate item–item rankings. This configuration captures coarse thematic groupings but intentionally excludes narrative detail and creative context.

As a result, the baseline tends to overgeneralize, grouping movies with shared high-frequency genres (e.g., *Drama*, *Action*) while failing to distinguish finer stylistic or narrative differences. Despite its limitations, this model serves as a necessary reference point for assessing the added value of richer metadata representations.

6.2 Enhanced Model: Combined Metadata TF-IDF

The enhanced model extends the baseline by incorporating multiple textual metadata fields, including genre, director, cast list, and plot overview, into a single representation prior to TF-IDF encoding. This richer document structure allows the model to capture both high-level thematic similarity and finer semantic cues related to creative style, narrative structure, and recurring collaborations.

By encoding this combined metadata using the same TF-IDF and cosine similarity framework, the enhanced model produces more differentiated and interpretable similarity scores while remaining directly comparable to the baseline.

6.3 Evaluation Strategy

In the absence of explicit relevance annotations, evaluation is conducted through structured qualitative analysis supplemented by quantitative summaries of similarity behavior.

Intuitive Alignment of Recommendations. For each query movie, we examine whether top-ranked recommendations exhibit coherent relationships in terms of narrative themes, creative contributors, or stylistic elements. Clear mismatches are treated as indicators of representational weakness.

Baseline vs. Enhanced Comparison. We compare the top–10 recommendations produced by the baseline and enhanced models for several representative query films, including *The Dark Knight*, *Inception*, and *The Shawshank Redemption*. This comparison highlights how additional metadata reduces genre-driven overgeneralization.

Qualitative Case Studies. Selected case studies examine similarity scores alongside contributing metadata, providing interpretability and insight into which features drive alignment in the enhanced representation.

6.4 Experimental Setup

All experiments use the same dataset, preprocessing pipeline, TF-IDF hyperparameters, and cosine similarity metric to ensure a controlled comparison. The only variation between models lies in the metadata fields included during vectorization. All results are reproducible using the accompanying notebook and application code.

7 Results and Analysis

We analyze MovieMate's behavior by inspecting recommendation outputs, examining similarity distributions, and comparing baseline and enhanced models. Because the task is unsupervised, results emphasize qualitative coherence and structural properties of the similarity space rather than accuracy against labeled ground truth.

7.1 Recommendation Examples

To illustrate system behavior, we examine representative recommendations produced by the enhanced model, which uses combined metadata for similarity estimation.

Example 1: *The Dark Knight*

- *Batman Begins* [29.8%]
- *The Dark Knight Rises* [29.3%]
- *The Prestige* [15.3%]
- *Brokeback Mountain* [12.9%]
- *Joker* [12.4%]

Example 2: *Inception*

- *Batman Begins* [13.0%]
- *Interstellar* [12.3%]
- *(500) Days of Summer* [12.1%]
- *Kagemusha* [10.5%]
- *Letters from Iwo Jima* [10.3%]

Example 3: *The Shawshank Redemption*

- *Mystic River* [10.9%]
- *Dev.D* [10.5%]
- *Lucky Number Slevin* [10.5%]
- *Pulp Fiction* [10.2%]
- *The Green Mile* [8.8%]

Across these examples, we observe that the enhanced representation often groups films based on shared directors, overlapping cast members, or closely related themes, producing recommendations that generally align with intuitive expectations.

Baseline Comparison. Compared to the enhanced model, the genre-only baseline frequently returns recommendations that share broad genre labels but lack narrative or stylistic coherence. For example, when querying *Inception*, the baseline prioritizes science-fiction and action titles without regard to director or narrative structure, whereas the enhanced model retrieves multiple Christopher Nolan films and thematically related works. This qualitative difference illustrates the benefit of incorporating richer linguistic metadata.

7.2 Quantitative Interpretation

Average Similarity Scores. To summarize similarity behavior, we compute the mean cosine similarity across the top–5 recommendations for a small set of query films.

Query Film	Avg. Similarity
The Dark Knight	0.2183
Inception	0.1201
The Shawshank Redemption	0.1059

Table 1: Average cosine similarity among the top-5 recommendations for selected query films.

Higher values indicate tighter clustering among recommended items. Films with strongly recurring stylistic signals—such as multiple works by the same director—tend to exhibit higher average similarity scores.

Recommendation Diversity. We also assess diversity by counting the number of unique genres represented in the top–5 recommendations for each query film.

Query Film	Unique Genres
The Dark Knight	6
Inception	8
The Shawshank Redemption	5

Table 2: Genre diversity among the top-5 recommendations for each query film.

These results suggest that the enhanced representation maintains thematic similarity while still retrieving films spanning multiple genres.

7.3 Qualitative Interpretation

Director Influence. We observe that films by the same director frequently appear together in the rankings. For example, multiple Christopher Nolan films are recommended for both *The Dark Knight* and *Inception*, reflecting the prominence of director information in the metadata.

Genre and Narrative Structure. For films such as *The Shawshank Redemption*, recommended titles tend to emphasize dramatic themes, moral conflict, or emotional intensity, indicating that narra-

tive structure and genre cues play a central role in similarity assessment.

Overall Behavior. Overall, the enhanced model produces more differentiated rankings than the genre-only baseline. By incorporating director, cast, and plot information, it reduces coarse genre-driven clustering and yields recommendations that are easier to interpret in terms of shared metadata.

8 Error Analysis

During evaluation, we observed several recurring error patterns attributable to the TF-IDF representation:

(1) Genre Dominance Bias. Because Drama is the most frequent genre in the dataset, similarity scores are often skewed toward drama-related associations. As a result, films from underrepresented genres (e.g., sci-fi comedy, Westerns, animation) occasionally receive drama-centric recommendations.

(2) Overreliance on Cast Names. Highly prolific actors contribute strongly to similarity scores, sometimes causing films with shared cast members to rank highly despite weak narrative overlap.

(3) Limited Plot Descriptions. Films with short or generic plot summaries provide insufficient lexical detail, reducing TF-IDF's ability to capture nuanced thematic similarity.

9 Limitations

Several limitations of MovieMate should be acknowledged:

- Lack of multimodal signals: The system relies exclusively on textual metadata and does not incorporate visual or audio information such as posters, trailers, or cinematography, which often encode important stylistic cues.

- No personalization: Recommendations are computed solely at the item–item level. The system does not adapt to individual user preferences or viewing histories.

- Dataset-restricted coverage: The IMDB Top-1000 dataset provides high-quality metadata but limited scope. Niche genres, international cinema, and low-budget films are underrepresented, which constrains recommendation diversity.

10 Ethical Considerations

Recommender systems can shape cultural exposure and viewing habits, raising concerns related to bias and diversity. MovieMate avoids many common ethical risks because it does not collect or rely on any personal user data. However, several considerations remain:

- Bias reinforcement: The dataset overrepresents Western and drama-oriented films, which may cause recommendations to reflect and amplify these existing biases.

- Reduced diversity exposure: As a content-based system, MovieMate favors stylistic similarity, potentially limiting exposure to unfamiliar genres, regions, or perspectives.

- Interpretability: A key ethical strength of MovieMate is transparency. Recommendations can be traced directly to shared metadata, allowing users to understand why a film was suggested.

11 Conclusion

In this project, we implemented MovieMate, a content-based movie recommendation system using TF-IDF vectorization over film metadata and cosine similarity for ranking. This approach enabled us to generate semantically meaningful recommendations without relying on user interaction data or large-scale model training.

Exploratory data analysis informed feature selection and revealed structural patterns—such as genre imbalance and director influence—that directly shaped recommendation behavior. While the system's simplicity imposes limitations, it also improves interpretability, scalability, and reproducibility.

MovieMate provides a practical baseline for content-based recommendation and establishes a clear foundation for more advanced extensions.

12 Future Work

Several directions could extend this work:

- Integrating collaborative filtering signals to capture user-specific preferences.

- Expanding metadata coverage through TMDB API integration to support a larger and more diverse catalog.

- Incorporating poster or trailer embeddings using CNNs or vision transformers.

- Adding temporal signals, such as recent popularity trends, to make recommendations more responsive to changing interests.

These extensions would improve personalization, thematic sensitivity, and overall recommendation quality.

A Appendix

This appendix lists foundational references related to content-based and collaborative recommender systems, information retrieval, and evaluation methodologies that informed the design and analysis of MovieMate.

(Lops et al., 2011) (Salton and Buckley, 1988) (Singhal, 2001) (Ricci et al., 2011) (Aggarwal, 2016) (Sparck Jones, 1972) (McNee et al., 2006) (Koren et al., 2009) (Resnick et al., 1994)

References

Charu C. Aggarwal. 2016. Recommender systems: The textbook. *Springer*.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. *Recommender Systems Handbook*, pages 73–105.

Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI Extended Abstracts*, pages 1097–1101.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 175–186. ACM.

Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. *Introduction to recommender systems handbook*. Springer.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Amit Singhal. 2001. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4):35–43.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.