Netflix Recommendation System

Content-Based Filtering with TF-IDF & SBERT

32,000 titles | 16K Movies + 16K TV Shows | 2010–2025

Data Overview

Exploratory Data Analysis

Dataset: 16,000 Movies + 16,000 TV Shows (TMDB, 2010–2025)

1,000 titles per year — balanced across time

397 overlapping show_ids between Movies & TV

Key features: genres, cast, director, description, country

Engagement signals: popularity, vote_count, vote_average

Data Quality Gaps

Feature	Movies	TV Shows
director	0.8%	68.5%
description	0.8%	20.0%
genres	0.7%	6.1%
cast	1.3%	7.2%

Genre taxonomy mismatch resolved: e.g. 'Sci-Fi & Fantasy' → 'Science Fiction, Fantasy'

System Architecture

Content-Based Filtering Pipeline

1. Data Processing
Merge Movies + TV, harmonize genres, handle missing values

2. Feature Engineering
TF-IDF soup (weighted concat of genres/cast/director/desc)

3. SBERT Embeddings
all-MiniLM-L6-v2 encodes tag soup into 384-dim vectors

4. Similarity
Cosine similarity across TF-IDF / SBERT / Hybrid matrices

5. Ranking
IMDB weighted rating + similarity → final_score (alpha=0.8)

6. App
Streamlit UI with search, genre filter, poster display via TMDB API

6 Approaches Compared

A. Combined (TF-IDF Soup)

B. Separate (TF-IDF Weighted Features)

C. Embedding — SBERT on Tag Soup

D. Embedding — SBERT on Description

E. Hybrid (TF-IDF + SBERT Tag)

F. Hybrid (TF-IDF + SBERT Desc)

Hypothesis Testing (H1–H3)

Statistical Validation of Design Decisions

H1: Multi-Feature > Genre-Only (Discriminative Power)

SIGNIFICANT ✓

`Wilcoxon W = 12,458 | p = 9.42e-51`

Multi-feature variance 6.8× higher (0.0009 vs 0.0001). Reject H_0.

H2: Feature Independence

SIGNIFICANT ✓

`Spearman ρ: genres–cast = 0.025, genres–desc = 0.093`

Features are effectively independent ($\rho < 0.2$). Weighted combination justified.

H3: Weighted Rating > Raw vote_average

SIGNIFICANT ✓

`Wilcoxon p = 1.04e-79 | Kendall τ = 0.064`

Weighted rating favors higher vote confidence (avg 1,237 vs 850 votes in top-5).

Hypothesis Testing (H4–H6)

TF-IDF vs SBERT Analysis

H4: Method Distinctness (TF-IDF vs SBERT)

`Mean Jaccard Similarity = 0.034`

Only 3.4% overlap in top-10 results. SBERT finds completely different items.

H5: Score Distribution (Loudness Problem)

`KS Statistic = 0.70 | p ≈ 0 | TF-IDF mean = 0.42, SBERT mean = 0.63`

SBERT scores are significantly higher. Normalization required before hybrid fusion.

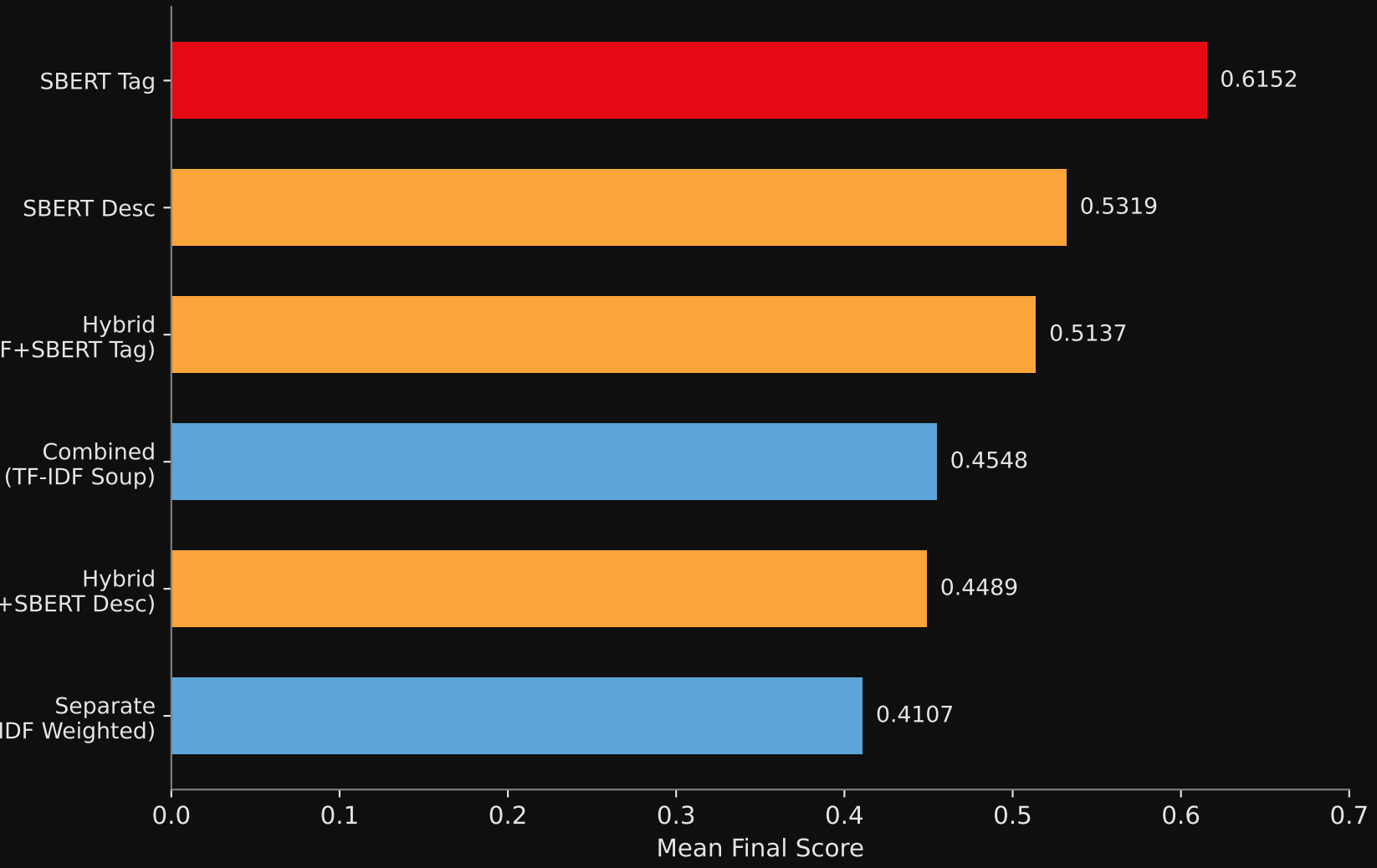
H6: Hybrid Quality Lift

`Wilcoxon p = 0.043 | TF-IDF rating = 5.82, Hybrid rating = 5.83`

Statistically significant but practically negligible improvement (+0.01 rating).

Approach Comparison

200-title benchmark — Mean Final Score



Winner: SBERT Tag

Mean Final Score: 0.6152

SBERT Tag outperforms all other approaches by a wide margin.

Hybrid approaches rank 3rd and 5th, not 1st as expected.

Why SBERT Outperforms Hybrid

Root Cause Analysis

1. The Loudness Problem (H5)

SBERT scores are ~50% higher than TF-IDF (mean 0.63 vs 0.42).

The 50/50 blend dilutes SBERT's strong signal, dragging hybrid score down.

2. Minimal Overlap = Signal Dilution (H4)

Only 3.4% Jaccard overlap — TF-IDF and SBERT find entirely different items.

Averaging two divergent rankings produces a compromise worse than either alone.

3. Tag Soup is Already Rich

Weighted tag (3×genre + 2×cast + director + desc) gives SBERT dense input.

SBERT captures latent themes/tone that TF-IDF bag-of-words misses entirely.

4. Naive Hybrid Weights

A fixed 50/50 split is suboptimal for this data.

Learned weights (e.g., 80/20 SBERT-favored) would avoid penalizing the stronger model.

Areas of Improvement

What Would Make This Better

Tune Hybrid Weights

Replace fixed 50/50 with grid search or learned weights (e.g., 80/20 SBERT-to-TF-IDF).

Handle Data Sparsity in TV

68.5% missing directors, 20% missing descriptions. Imputation or fallback strategies needed.

Upgrade Embedding Model

all-MiniLM-L6-v2 (384-dim) is lightweight. Larger models like all-mpnet-base-v2 could help.

Add Collaborative Filtering

Purely content-based now. User interaction data would enable content + collaborative hybrid.

Evaluation with Human Judgement

Current metric is automated. A/B testing or user studies would validate real-world relevance.