Netflix Content Classification and Recommendation System

Team ID: 06

Team Members: J Joshua Haniel (RA2311003040056)
S Yashwant (RA2311003040116)

Project Guide: Dr. G.Paavai Anand

INTRODUCTION AND MOTIVATION

- OTT platforms like Netflix have a massive and growing content library.
- Users struggle to find relevant content quickly.
- Classification of titles (Movies vs. TV Shows) and recommendation engines can improve user experience.
- Motivation: Enhance content discovery using AI/ML.

SCOPE OF THE PROJECT

- Automated classification of Netflix titles based on descriptions.
- Content-based recommendation system to suggest similar shows/movies.
- Support for both keyword-based (TF-IDF) and semantic (transformer embeddings) recommendations.
- Interactive UI using Gradio for easy exploration.

ABSTRACT

This project presents a machine learning–driven system to classify Netflix titles and recommend similar content. Logistic Regression, Naive Bayes, and Support Vector Machines were evaluated on TF-IDF features to predict whether a title is a Movie or a TV Show. Additionally, two content-based recommendation engines were built—one leveraging TF-IDF similarity and another using sentence-transformer embeddings for semantic similarity. The system is deployed with an interactive UI to demonstrate real-world usability.

LITERATURE SURVEY

- Content-Based Filtering: Prior works rely on metadata similarity (e.g., TF-IDF of movie plots).
- Collaborative Filtering: Requires user-rating data, often unavailable for proprietary datasets.
- Deep Learning for NLP: Transformers (like BERT, MiniLM) achieve better contextual understanding for recommendations.

EXISTING METHODS

Category	***Existing***	***Description***	***Advantages / Limitations***
Existing	Manual Browsing & Search	Users manually browse categories, search by keywords, or rely on Netflix's in-built UI.	Limitations: Time-consuming, not personalized, difficult to scale with growing content.
Existing	Rule-based Metadata Filtering	Filters based on genre, year, or country metadata.	Limitations: Rigid, ignores context and semantics, often gives irrelevant results.
Existing	Basic Content-Based Filtering (TF-IDF only)	Uses simple TF-IDF on text (like plot/description) to compute similarity.	Limitations: Ignores deep semantic meaning, requires large feature space, only moderately accurate.
Proposed	Classical ML with Advanced TF-IDF Models	Logistic Regression, SVM, Naive Bayes on TF-IDF features for classification (Movie vs TV Show).	Advantages: Strong baseline, interpretable, good accuracy. Limitation: Context understanding is still shallow.
Proposed	Transformer-based Semantic Recommendation (Sentence-BERT / MiniLM)	Generates sentence embeddings of descriptions and metadata, computes similarity for recommendations.	Advantages: High accuracy, semantic understanding, scalable. Limitation: Requires more computational resources.
Proposed	Interactive Deployment via Gradio UI	Unified interface with tabs for classification and recommendation (TF-IDF & Semantic).	Advantages: Easy to use, interactive, presentation-ready. Limitation: Needs stable environment for deployment.

PROBLEM STATEMENT

Users on OTT platforms face difficulty in finding relevant content due to:

- **Huge volume of titles.**
- **Limited traditional recommendation systems.**
- **Lack of intelligent classification of new content.**
-

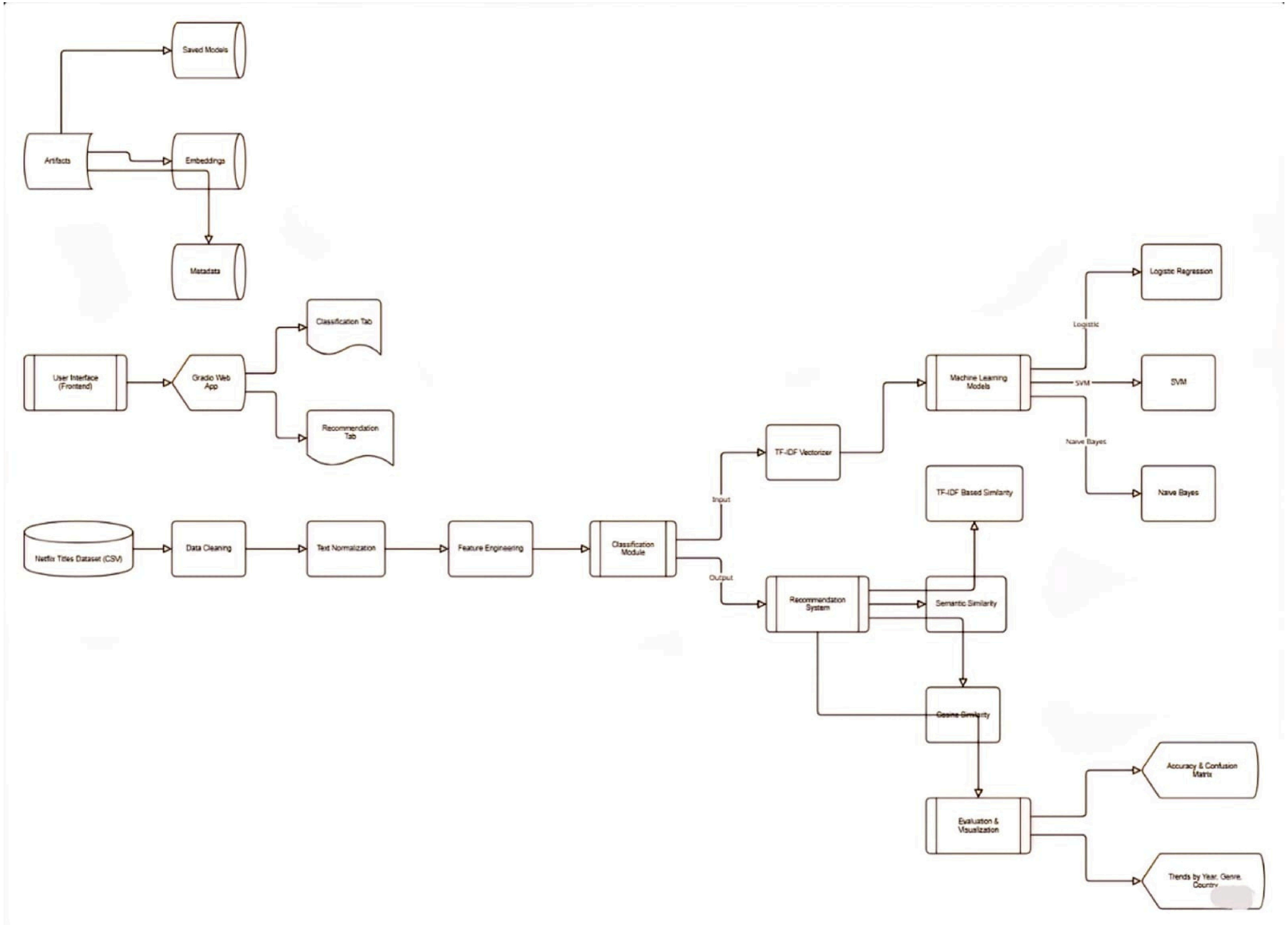
Our problem statement:

"Can we build an ML-based system to classify Netflix titles and provide intelligent, accurate content recommendations?"

OBJECTIVES

1. Build an ML model to classify titles as Movies or TV Shows.
2. Compare multiple models (Logistic Regression, SVM, Naive Bayes).
3. Develop content-based recommenders using TF-IDF and semantic embeddings.
4. Provide an interactive UI for classification and recommendations.

ARCHITECTURE DIAGRAM



LIST OF MODULES

1. Data Preprocessing
2. Classification (Movie vs. TV Show)
3. TF-IDF Based Recommendation
4. Semantic (Transformer) Recommendation
5. Visualization & Evaluation
6. Gradio UI Deployment

MODULE DESCRIPTION

- Data Preprocessing: Cleaning, normalization, handling missing values.
- Classification: TF-IDF + ML models, tuned via GridSearchCV.
- TF-IDF Recommendation: Keyword overlap–based similarity.
- Semantic Recommendation: Sentence embeddings for deep semantic matching.
- Visualization: Trends by year, genres, countries, confusion matrices.
- UI Deployment: Interactive Gradio interface with recommendation and classification tabs.

REFERENCE

1. Suvarna, V. (2023). Netflix Titles Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/venkateshsuvarna27/netflix-title
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
3. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.
4. Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433–459.
5. Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to Recommender Systems Handbook. Springer.
6. Abhishek, T. (2020). Building Recommendation Systems with Machine Learning and AI. Packt Publishing.
7. Gradio. (2023). Gradio Documentation. Retrieved from https://www.gradio.app

THANK YOU