Anime Recommendation system using Content-Based Filtering (CBF) and Sentiment Analysis Hybrid

Kapil Singh Negi
SoCS, UPES, Dehradun, Uttarakhand, India
kapil.negi12300@gmail.com

ABSTRACT

Purpose – The global anime market was valued at over $24 billion in 2021 and is projected to reach over $60 billion by 2030. With its wide range of genres and unique visual characters, anime is a popular form of animation with a large global audience. Tens of thousands of anime have been produced, and it is simple for an anime to go unnoticed. The majority of anime fans mainly watch anime based on recommendations from friends, which is hardly efficient; most viewers become biased, and many excellent anime are missed by the intended audience. This paper proposes an AI anime recommendation system that addresses this very problem. It minimizes the bias introduced by the majority using content-based filtering (CBF) and concentrates more on the context and theme of anime to better suit the needs of the user, allowing him to find anime that suits his taste rather than being forced to watch the popular series.

Research limitations/implications The research is limited by the scope of the dataset used and the system is severely hampered when dealing with new users (no history). The vast, mostly empty user-item matrix challenges the training of the Deep Collaborative Filtering (CF) model, especially for niche recommendations. Explicit feedback is non-neutral, as highly engaged users often only rate extreme preferences (love/hate), introducing selection bias into the training data.

I. INTRODUCTION

The art, animation techniques, production, and process of anime are different from those of other animation genres. Anime displays a diverse range of visual styles that vary throughout creators, artists, and companies. Although anime is not dominated by any one art form, there are certain similarities among them in terms of character design and animation methods.

Children's, girls', boys', young men's, young women's, and a wide variety of genres aimed at an adult audience are among the many target demographics used to categorize anime. In an effort to appeal to a wider audience, shōjo and shōnen anime occasionally include features that are well-liked by kids of all genders. In addition to adult themes and events, adult anime may have a slower pacing or more intricate plots that younger viewers could find uninteresting.

Over 200 animation works are aired in Japan each year, and the industry for animation as a whole, including associated products, has grown to be worth about 1.8 trillion yen in recent years. Anime is a popular form of animation with a large global audience with audience from many age groups and diverse cultures. Along with the audience the anime has also evolved from just being influenced by Japanese to many other cultures. Some international successes include *Sailor Moon* and *Dragon Ball Z*, both of which were dubbed into more than a dozen languages worldwide.

With the growing supply and demand in the anime it is only getting difficult for anime fans to find the anime which suits their interest, for the majority the type of anime they watch is highly influenced by their peer group as an anime fan can only go reliably to his friends to get a good recommendation. But this method is imperfect, as most times it just so happens that one person 'person A' likes some anime and wishes to know if there are more anime like this, his friend 'person B' may not know of more anime like that because the type he 'person B' watches are of a different genres, animation style or differently paced in such cases anime fan 'person A' needs to scrape the internet starting reddit comment chain, joining discords and what not.

With this research I aim to significantly ease the burden on people in similar situation to 'person A' including myself using a artificially intelligent system that can identify the type of content 'person A' enjoys watching and recommend me similar content.

In this paper I will be using Content based filter on the database [https://www.kaggle.com/datasets/dbdmobile/myanimelist-dataset]

For the sake of simplicity, I will be using only anime-dataset-2023.csv file within the database. The synopsis data will be used for sentiment analysis of the anime, and the rest of features will be used for Content-Based filtering.

For Sentiment analysis, TF-IDF will be the preferred method to convert textual data into numeric weights, but I will also show the comparison with Bag of words method, as for the prediction model the primary focus will be on Logistic regression and will further compare with some other classifiers.

For Content-Based filtering the database will be optimised using label encoding and one hot encoding for categorical data in features. The Cosine similarity will be highly efficient with the current database.

For the fusion of the scores from both methods I will be using a weighted scoring Then to fine tune the model by optimizing the weight, (alpha)

The advantages of hybrid model are-

1. It eliminates the Cold Start Issue for Items: Using CBF, the model can still suggest an anime based on its features even if it has no ratings.
2. It refines similarity: two Action/Adventure shows may have structural similarities, but the SA score distinguishes between the whimsical, lighthearted one and the gritty, dark one.

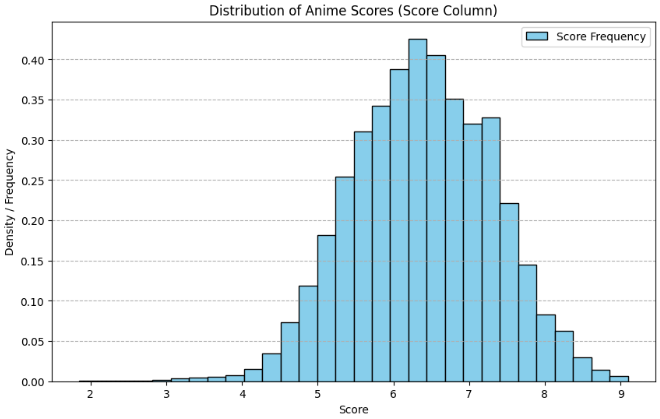
II. METHODS AND MATERIAL

Here we discuss the detailed approach that we have followed to process the database and create the model.

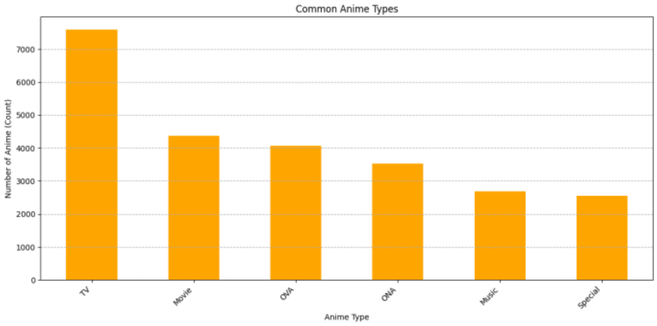
Dataset: I will be using a free dataset from Kaggle, [https://www.kaggle.com/datasets/dbdmobile/myanimelist-dataset],
this dataset is under Open Database License (ODbL) 1.0 [http://opendatacommons.org/licenses/dbcl/1.0/].
Within this data set I will be particularly be using the anime-dataset-2023.csv for data extraction.

Visualization: The database contains the following features-

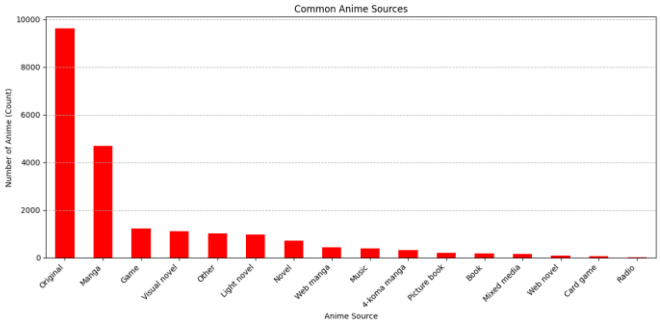
- **anime_id**: Unique ID for each anime.
- **Name**: The name of the anime in its original language.
- **English name**: The English name of the anime.
- **Other name**: Native name or title of the anime(can be in Japanese, Chinese or Korean).
- **Score**: The score or rating given to the anime.
- **Genres**: The genres of the anime, separated by commas.
- **Synopsis**: A brief description or summary of the anime's plot.
- **Type**: The type of the anime (e.g., TV series, movie, OVA, etc.).
- **b**: The number of episodes in the anime.
- **Aired**: The dates when the anime was aired.
- **Premiered**: The season and year when the anime premiered.
- **Status**: The status of the anime (e.g., Finished Airing, Currently Airing, etc.).
- **Producers**: The production companies or producers of the anime.
- **Licensors**: The licensors of the anime (e.g., streaming platforms).
- **Studios**: The animation studios that worked on the anime.
- **Source**: The source material of the anime (e.g., manga, light novel, original).
- **Duration**: The duration of each episode.
- **Rating**: The age rating of the anime.
- **Rank**: The rank of the anime based on popularity or other criteria.
- **Popularity**: The popularity rank of the anime.
- **Favorites**: The number of times the anime was marked as a favorite by users.
- **Scored By**: The number of users who scored the anime.
- **Members**: The number of members who have added the anime to their list on the platform.
- **Image URL**: The URL of the anime's image or poster.



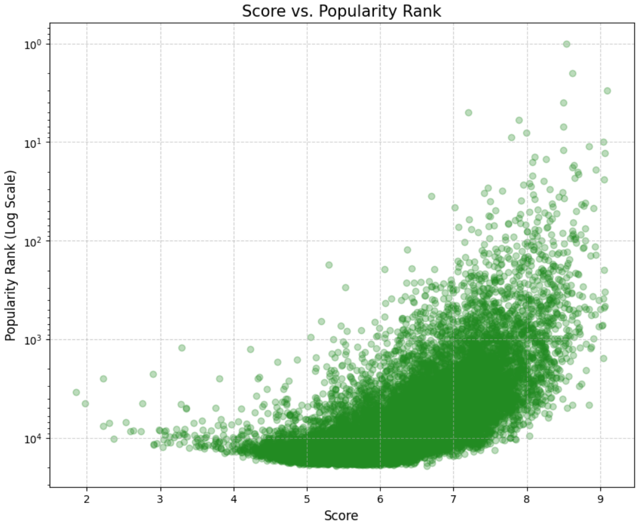
Distribution of anime scores in database



Most common anime types in the database



Most common anime sources



Score vs Popularity

To process the database, we have to perform two things first to process the synopsis column using TF-IDF to data into numeric weights by

1. Lowercasing and Punctuation Removal
2. Tokenization and Stop Word Removal
3. Lemmatization

Then converting this into TF-IDF,

1.Calculate Term Frequency (TF) for each word in each document

2) Calculate the Inverse Document Frequency (IDF) for each word

3) Multiply the TF and IDF values for each word to get its TF-IDF score

For the rest of the features, we will use label encoding and one hot encoding to optimize the database for machine learning.

The database will be divided into 60% testing data and 40% testing data since I am working with relatively small amount of data

Then to apply logistic regression on the synopsis data, and cosine similarity to other features and combining them via weighted scoring to calculate a hybrid score

$$\text{Hybrid Score} = (\alpha \times \text{CBF Score}) + ((1 - \alpha) \times \text{SA Score})$$

This hybrid score will be needed to be fine-tuned manually.

Finally, I will be comparing this hybrid model with baseline-

Pure CBF: Using *only* the metadata columns.

Pure SA: Using *only* the sentiment score for similarity.

III. REFERENCES

[1] W. K. Tan and R. M. F. R. M. N. Rashid, “Intelligent Mobile Agents-Based System for Healthcare Monitoring,” in *Advances in Computing and Data Sciences*, S. Singh, A.K. Kar, V.K. Singh, and N.S. Raghuwanshi, Eds. Cham, Switzerland: Springer.

[2] Puji Lestari. 2012. *A New Concept in Developing English-Indonesian Digital Dictionary with Semantic Network Model*. (Undergraduate Thesis, UIN Maulana Malik Ibrahim, Malang, Indonesia).

[3] dbdmobile. 2025. *MyAnimeList Dataset*. (Accessed Oct. 2025), Available: https://www.kaggle.com/datasets/dbdmobile/myanimelist-dataset

[4] *Anime*. (Accessed Oct. 2025), Available: https://en.wikipedia.org/wiki/Anime#Markets