AniMatch: A Content-Based Anime Recommendation System

Berkay Orhan Jonatan Ebenholm

Abstract

AniMatch is a content-based recommendation system designed to suggest anime titles based on their inherent features. The system employs BERT embeddings to extract meaningful representations of anime metadata and cosine similarity to compute relationships between titles. Unlike traditional recommendation systems, AniMatch focuses on metadata analysis, ensuring accurate and contextually relevant recommendations.

1 Introduction

Recommendation systems have become indispensable in digital platforms for personalizing user experiences. Traditional systems often rely on user data, raising concerns about privacy and usability in cold-start scenarios. AniMatch addresses these challenges by adopting a content-based approach, focusing on the intrinsic features of anime metadata to deliver recommendations.

2 Background

Content-based recommendation systems utilize the attributes of items to identify similarities and generate recommendations. In the context of anime, metadata such as descriptions, genres, themes, and demographics serve as the foundation for such systems. This type of system, in contrast to a collaborative-filtering system, works perfectly for systems which have no user data, Since the website which we made does not have a login system or cookies attached. We opted out of doing collaborative-filtering since the website is meant to be accessed, used and left in a span of a few minutes, in which an account setup would take longer than the use of the site itself, leading to what the we believe to be a worse user experience. With our need and the advancements in natural language processing, models like BERT have enabled the extraction of deeper semantic meanings, making them ideal for metadata-based recommendations, and this is what was used in the project.

3 Purpose and Research Question

The purpose of this study is to explore the effectiveness of content-based filtering using BERT embeddings in anime recommendation systems. The research question is: *How effective is a content-based approach utilizing BERT embeddings and cosine similarity in generating relevant anime recommendations?*

4 Data Collection and Preparation

4.1 Generality of Data

The datasets used in AniMatch is the *Top Anime Dataset 2024*, sourced from Kaggle, and *Jikan API*, sourced from Jikan. The data from Kaggle contains the features shown in the list below for 1,000 anime titles.

1. **Score**: The rating or score assigned to each anime title.
2. **Popularity**: Measure of how popular each anime is among viewers.
3. **Rank**: Ranking of each anime title within dataset.
4. **Members**: The number of members or viewers associated with each anime.
5. **Description**: A brief overview or summary of the plot and themes of each anime.
6. **Synonyms**: Alternative titles or synonyms used for each anime.
7. **Japanese Title**: Original title of the anime in Japanese.
8. **English Title**: English-translated title of the anime.
9. **Type**: Classification of anime type (e.g. TV series, movie, OVA, etc.).
10. **Eps**: Total number of episodes in each anime series.
11. **Status**: Current status of the anime (e.g., ongoing, completed, etc.).
12. **Aired**: Date range of when the anime was aired.
13. **Premiered**: Date when the anime premiered for the first time.
14. **Broadcast**: Information about the broadcasting platform or channel.
15. **Producers**: Companies or studios responsible for producing the anime.
16. **Licensors**: Organizations or companies holding the licensing rights for the anime.
17. **Studios**: Animation studios responsible for producing the anime.
18. **Source**: Original source material for the anime (e.g., manga, novel, original).
19. **Genres**: Categories or genres that the anime belongs to.
20. **Demographic**: Target demographic audience for the anime (e.g., shounen, sjoujo, seinen, josei).
21. **Duration**: Duration of each episode or movie.
22. **Rating**: Content rating assigned to each anime (e.g., G, PH, PH-13, R).

The Jikan data is much the same, but it also has two extra features that were useful for the project, those being:

- Themes: Categories that are not genres, but still describe the setting of the world. As an example, think military or gore.
- Images: The poster image of the anime.

4.2 Source and Selection

Most of the features that were scraped from Kaggle are not considered relevant for an accurate recommendation. Without going into any detail about why a feature was not chosen, instead we will explain which were chosen and why.

First and foremost, number five on the list above, **Description**, was used since a similar anime should be able to be described in similar terms. If the description explains a show to be bloody yet adventurous, another show with that description is probably a good match.

Secondly, number nineteen, **Genre** was used since even if an anime might not share a description, if a user likes a show based on its *sci-fi* or *fantasy* elements, then it makes sense to recommend another anime based on that.

Third, number twenty two on the list, the **Rating**, was used to find similarities. This might be controversial, but aims to tackle the issue that some genres are not age specific. As an example, *The lord of the rings* movie trilogy is a fine trilogy to watch with children of appropriate age, but if the age rating of a recommender system looked at the genres of the trilogy and matched the fantasy genre to *Game of Thrones*, then it would have found a show that does indeed fit the genre, but is a horrible match in how explicit the latter example is. Therefore, an age rating is useful when finding similar animes, although the example given of course are not animes.

Lastly we got the demographic, number twenty on the list, to determine if a show was meant for a similar audience. This is useful since especially anime that is meant for teenagers are usually meant for either boys or girls with is what shounen and shoujo is meant to represent respectively.

The two additional features from Jikan helped with different purposes. Themes is also used to help with recommendation, since if one show could be rated R but not contain gore or other similar themes. Therefore, being able to compare the themes of a show is good to see if they are somewhat similar.

The images from Jikan were used to attach an image to the anime on the website, both because it made the website look pretty, but also because it helps the user associate the anime to an image.

4.3 Data Processing and Cleaning

Data preparation included:

1. Removal of missing or incomplete entries.
2. Conversion of text fields to lowercase for uniformity.
3. Removal of stop words, special characters, and redundant spaces.
4. Tokenization of text fields for embedding generation.

4.4 Normalization

All numerical features were normalized using min-max scaling to ensure uniformity across data points. Text data was preprocessed for BERT embeddings.

5 Modeling

5.1 Recommender System

AniMatch employs a content-based recommendation system. Each anime is represented by its metadata, which is encoded into numerical embeddings using BERT.

5.2 Cosine Similarity

Cosine similarity is used to compute the closeness of two anime embeddings. It is defined as:

$$\cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|} \tag{1}$$

where $\vec{A}$ and $\vec{B}$ are the embedding vectors for two anime titles. Higher similarity scores indicate closer matches.

5.3 Implementation

The system was implemented in Python mainly using the following libraries:

- **Transformers**: For generating BERT embeddings.
- **scikit-learn**: For computing cosine similarity.
- **pandas** and **numpy**: For data preprocessing and handling.

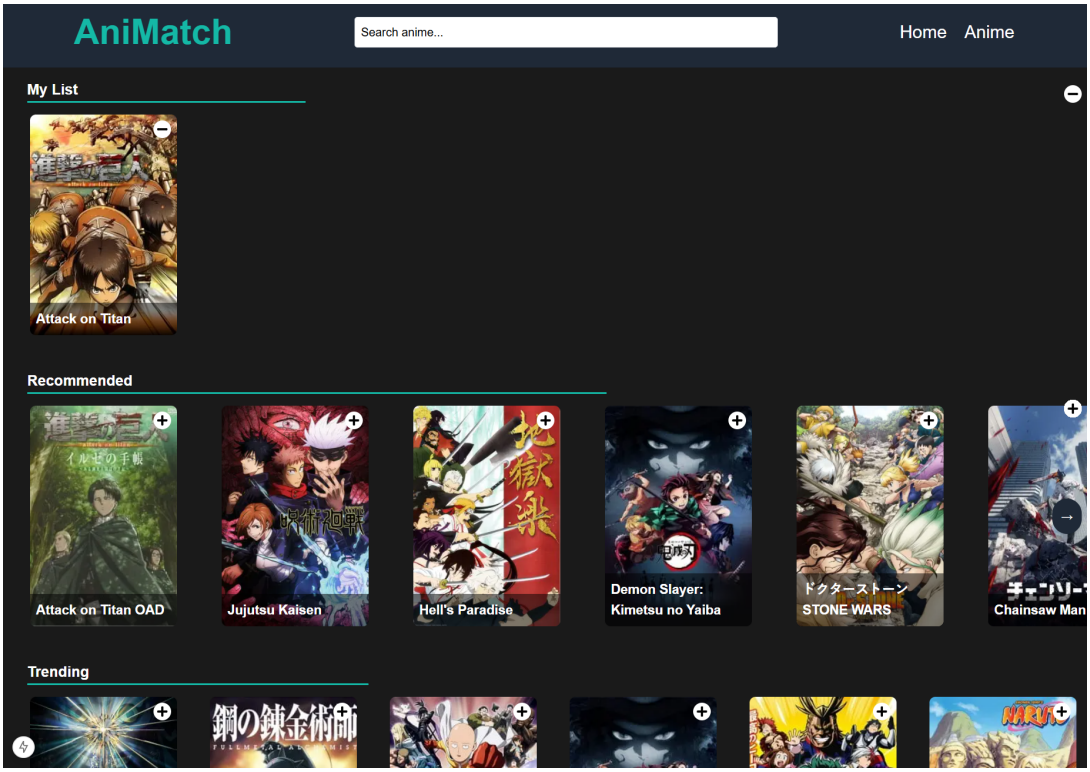

Figure 1: Frontpage of the website.

6 Evaluation

6.1 Description of Model

The recommendation model generates a ranked list of the top similar anime for a given title. This ranking is based on cosine similarity scores. The result of this can be seen 1.

6.2 Relevance and Efficiency

The system provides accurate and contextually meaningful recommendations .Precomputing embeddings significantly reduces query time, ensuring fast response rates.

6.3 Example Results

For the input *Attack on Titan*, the recommendation system assigns equal weight to each feature embedding category. The weight distribution amongst the different features is as follows:

Feature	Weight
Themes	0.25
Age rating	0.1
Description	0.25
Demographic	0.15
Genre	0.25

Each feature is assigned a weight of 0.2, collectively summing to 1.0. With the given weight distribution, the system recommended the following:

1. **Demon Slayer**: A high-paced action anime with themes of family bonds and overcoming challenges, much like *Attack on Titan*'s focus on humanity's survival and camaraderie.

2. **Jujutsu Kaisen**: Parallels AoT's high-stakes battles against existential threats (curses vs. Titans) and explores the psychological toll of power. Characters like Yuji and Eren bear the burden of harboring destructive forces, balancing duty with personal trauma in a world teetering on annihilation.

3. **Hell's Paradise**: Shares *Attack on Titan*'s bleak survivalist ethos, pitting flawed humans against nightmarish creatures in a fight for existence. Themes of redemption, the cost of freedom, and blurred lines between humanity and monstrosity resonate deeply with AoT's moral complexity.

4. **Dr. Stone: Stone Wars**: Like *Attack on Titan*, this series pits humanity against existential collapse, emphasizing survival through strategy and unity. Both explore rebuilding civilization amid relentless threats, though *Dr. Stone* replaces Titans with primal human conflict and scientific ingenuity.

5. **Chainsaw Man**: Mirrors *Attack on Titan*'s visceral tone and themes of sacrifice, moral ambiguity, and grotesque adversaries. Protagonists Denji and Eren grapple with identity loss and monstrous transformations to protect others, while confronting cycles of violence and exploitation.

Attack on Titan is a critically acclaimed anime set in a dystopian world where humanity faces extinction from giant humanoid creatures known as Titans. The story follows Eren Yeager and his friends as they join the military to fight these Titans, uncovering dark secrets about their world along the way. Its themes of survival, sacrifice, and the blurred lines between good and evil make it a complex and emotionally gripping series.

The recommendations align with common narrative and thematic preferences of *Attack on Titan* fans, providing a mix of action, emotional depth, and survival themes. This demonstrates how AniMatch effectively captures and recommends contextually relevant titles.

6.4 Other results

Other than the recommender system itself, a locally hosted website was made as well. We would of course like to highlight some of it as well. Firstly, we made a search feature. This allows a user to easily look up an anime. It can be seen in 2. Secondly, we implemented some generic scrolling categories. The categories currently are only the trending anime, which is more like the most voted on anime of all time, and also the top ranked anime, which is like it sounds the highest rated animes. They can be seen in 3. Lastly, if the user would like to read a bit more about an anime, they can press it, either in the search bar or in the scrolling categories, and they will be forwarded to the anime description page which has all the information about the anime that a user would need. One such page can be seen in 4. The description page also shows reviews and recommendations based on it.

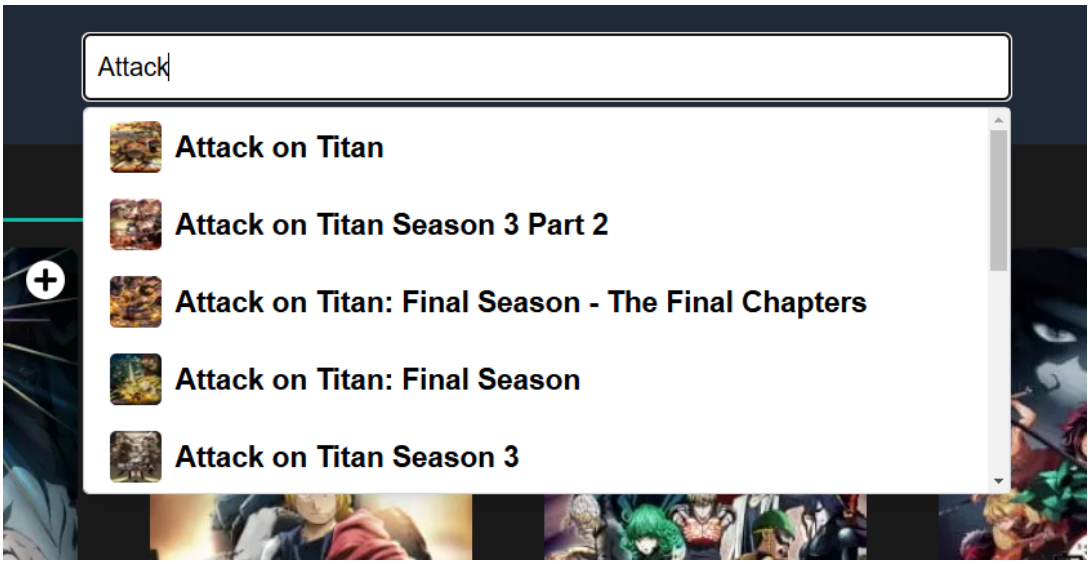

Figure 2: Search bar on the navigation bar.

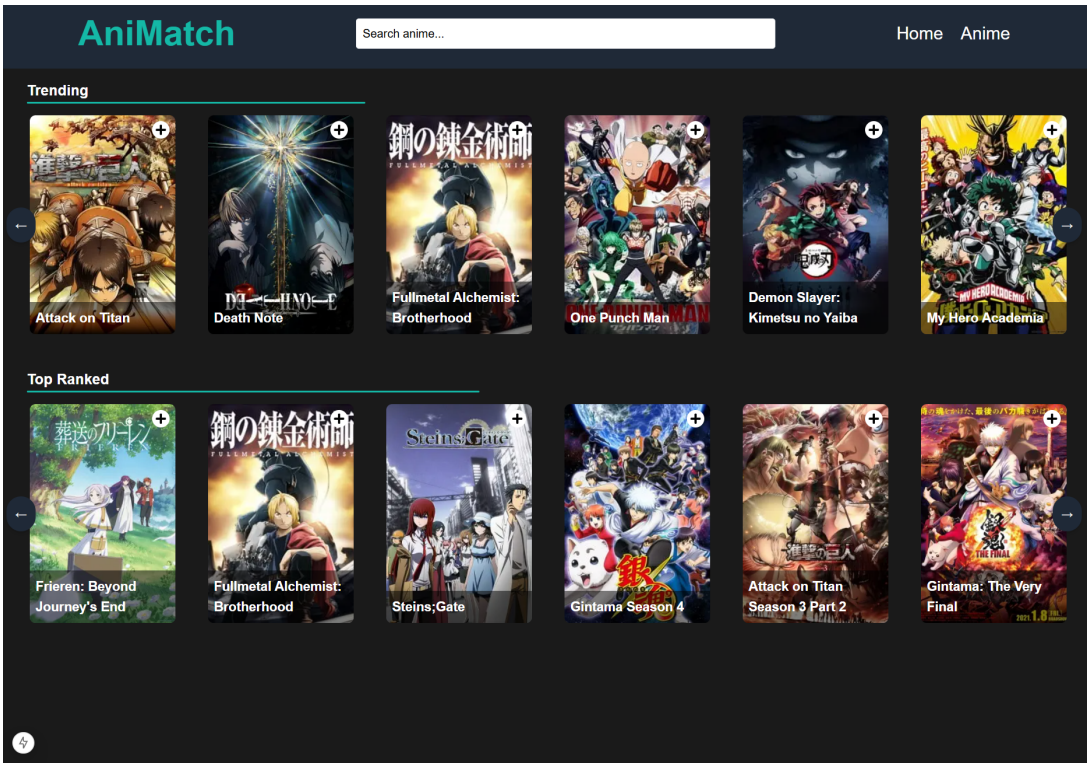

Figure 3: Front-page categories.

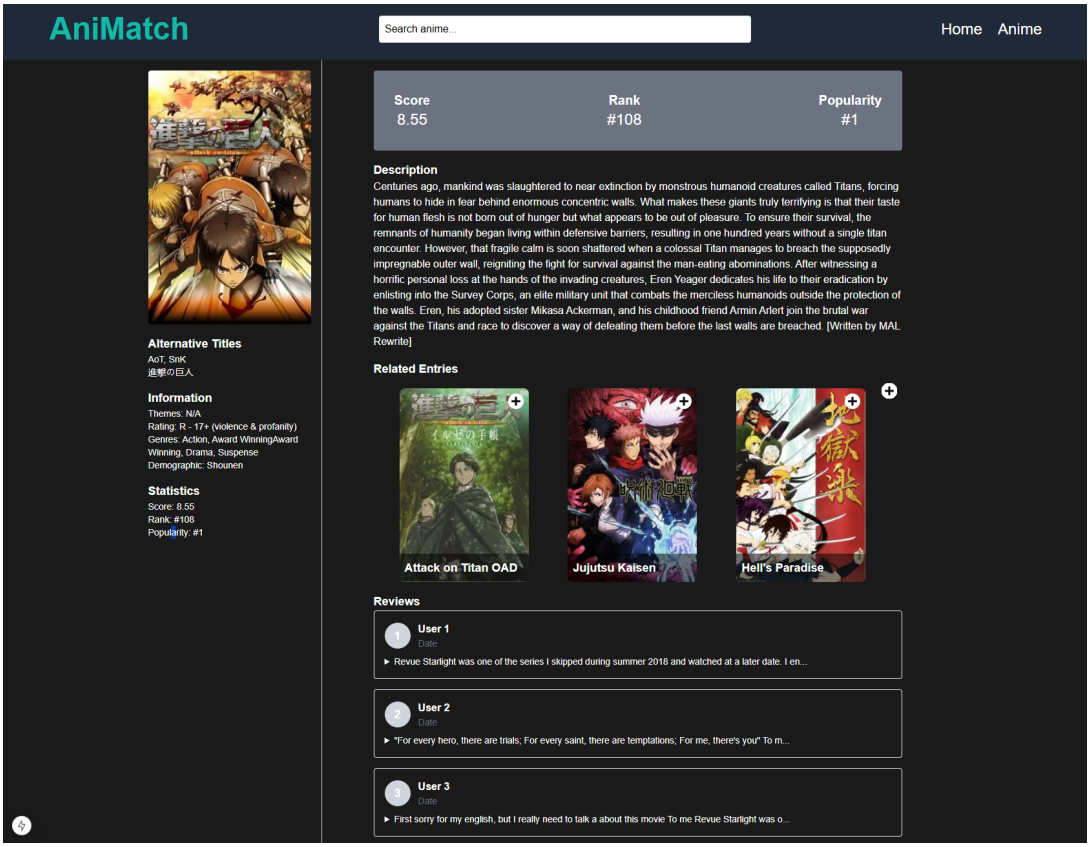

Figure 4: Description page of Attack on titan.

7 Limitations

7.1 Feature Dependencies

The system relies heavily on the quality and diversity of metadata. Missing or biased metadata can affect recommendation accuracy.

7.2 Cold-Start Problem

While the content-based approach avoids reliance on user data, it may struggle with newly added anime that lack sufficient metadata.

7.3 Serendipity

Currently the user would experience a lot of serendipity, since sometimes there are not a lot of similar matches to an anime. This probably has less to do with the system being robust, and more with the data set being relatively small as compared to the number of features, which is to say that there are a lot of possible feature-vectors to compare with relative to the amount of anime. This means that there are not too many anime that will be a super close match to the current input of the user, leading to anime that is somewhat similar but not very to fill in the gaps.

7.4 Reinforcement loop

There is also the issue of reinforcement looping. If the user picks an anime like attack on titan, and then they watch a recommended anime based on the recommender system output, if they

then were to add that show to the input of the system and ask for a recommendation based on it, the system would spit out the same shows as it did for attack on titan. This is because we are comparing the angles between the features of the animes, but since the angles in this case are between the recommended anime and attack on titan, asking for another recommendation based on those two would again yield similar angles and therefore similar results. The way to break this reinforcement loop currently is to add another anime that is not a super close match and hope that it has anime that is more similar to it than the other anime in the input list.

8 Discussion

8.1 Analysis of Results

The results highlight the effectiveness of BERT embeddings in capturing semantic relationships. However, the computational cost of generating embeddings remains a challenge, but luckily these embeddings only have to be generated once and can be later on stored in a database for use. Reinforcement loops are also a big limitation with the system, as it would not be okay in a streaming service like Netflix, that has users expecting new and interesting shows to watch, for the system to stop recommending new shows and instead only recommend what they have already watched.

8.2 Alternative Methods

Alternative approaches, such as hybrid systems combining collaborative and content-based filtering, could enhance personalization. Additionally, fine-tuning BERT for anime-specific metadata might improve accuracy.

8.3 Suggestions for Improvement

- Integrate user preferences to create a hybrid recommendation system.
- Explore lightweight embedding models to reduce computational costs.
- Fine-tune BERT embeddings for anime-specific contexts.

9 Conclusion

9.1 Summary of Findings

AniMatch demonstrates the potential of content-based filtering using BERT embeddings and cosine similarity in anime recommendation systems. The model provides accurate and relevant recommendations while ensuring privacy.

9.2 Reconnection to Research Question

This study demonstrates that a content-based approach utilizing BERT embeddings and cosine similarity is highly effective in generating relevant anime recommendations. The use of embeddings captures nuanced context from metadata, and computational costs are minimal since embedding generation is a one-time preprocessing task. The results validate the approach's efficacy in meeting the study's goals of relevance and efficiency.

References

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

2. Kaggle. (2024). Top Anime Dataset 2024. Retrieved from `https://www.kaggle.com/datasets/bhavyadhingra00020/top-anime-dataset-2024`.

3. Jikan. (2024).

4. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.