Emotional Aware Anime Recommendation System using NLP based Dialogue Analysis

---

Abstract

Aims/Objectives: Traditional anime recommendation systems frequently rely on viewing history, user ratings, and genre-based filtering, all of which often fail to capture the complex emotional immersion inherent in the medium. Given that emotional resonance is a primary driver of audience engagement, incorporating affective data can significantly enhance recommendation accuracy.

Study Design: Quantitative experimental research with hybrid machine learning framework.

Methodology: This study proposes an emotion-aware hybrid anime recommendation system that extracts emotional patterns from dialogue using Natural Language Processing (NLP). The methodology involves preprocessing over 1.5 million anime subtitle lines to eliminate noise and segmenting dialogue for fine-grained analysis. A transformer-based emotion detection algorithm is applied to each dialogue line to generate probability distributions across seven emotion categories. These forecasts are aggregated to produce episode-level emotion vectors, which are further augmented with engineered "anchor features"—Total Intensity, Emotion Breadth, and a Positivity Index—to ensure architectural resilience against subjective user reporting.

Results: Within a hybrid framework, user emotional preferences are compared with anime profiles using similarity-based scoring and semantic embeddings. Experimental results confirm that the integration of these 9D features produces a 33% to 65% improvement in noise resilience across all architectures, with XGBoost emerging as the optimal backbone for real-world deployment.

Conclusion: This approach yields suggestions that are more context-aware, psychologically grounded, and aligned with the emotional expectations of viewers.

Keywords: Emotion-aware recommendation systems; NLP; Subtitle-based emotion analysis; Transformer-based detection; Episode-level emotion modeling; Hybrid recommendation framework; Noise resilience; Anchor features.

2010 ACM Computing Classification System: Information systems → Recommender systems; Computing methodologies → Natural language processing; Computing methodologies → Supervised learning.

1 Introduction

1.1 Background of the Research

In the modern era there is an upward trend toward anime as it is becoming a peak entertainment medium among audiences. As anime series heavily emphasise emotional immersion through both dialogues and visual effects, many anime characters remain deeply embedded in the audience's memory over long periods. Because of this, the targeted audience of anime lovers is increasing day by day. The anime industry has a considerable impact on both the economic and cultural sectors, which can be clearly identified through its expected global market expansion. By 2034, the global anime market—which was estimated to be worth $81.96 billion in 2024—is predicted to have more than doubled, expanding at a rate of 9.53% to reach over $203.68 billion [1]. Furthermore, Netflix, a worldwide streaming platform, stated that half of its users actively watch anime [2]. This expansion of the anime industry necessitates recommendation systems that prioritise personalised preferences not only based on genre and ratings but also on the user's mental state, as audiences expect a degree of emotional resonance and even therapeutic value from what they watch. Traditional recommendation systems have failed to capture this specific user need; as a result, anime within the same genre can have very different emotional profiles.

According to recent research, emotion-aware recommendation systems address “individual emotional states and preferences,” seeking to capture both explicit preferences and the subtle emotions people display with respect to films [3]. This depicts the emerging trend toward focusing on emotional states and proves the current tendency towards analysing emotions. Transformer-based language models can help bridge this gap, as they are capable of extracting emotional characteristics and identifying linguistic relationships [4]. According to recent research, transformer architectures outperform traditional machine learning techniques in recognising nuanced emotional states and comprehending contextual emotions in conversation [5, 6]. A study conducted in 2010 explored recognition of emotional speech patterns through four types of emotions—joy, sadness, anger, and surprise—and also supports the emotional prioritisation approach [7]. Despite these advances, integrating dialogue-level emotional analysis into a recommendation pipeline has remained largely unexplored.

This research aims to propose a subtitle-based, emotion-driven anime recommendation framework for episode-level emotional classification. This personalised approach is primarily aimed at offering a more human-centred and psychologically grounded recommendation experience for users.

1.2 Research Problem

At present there are several kinds of anime recommendation systems based on genre, ratings, or user feedback. There is a clear lack of a user-centred, emotion-based approach. From a research perspective, these factors are insufficient for successfully extracting the emotional level of anime. There are two predominant approaches in machine learning for recommendation systems: collaborative filtering and content-based filtering. Collaborative filtering faces limitations such as new-user cold-start constraints and repetitive recommendations, while content-based filtering over-relies on metadata, tags, and labels. Because of these existing limitations, these approaches fail to successfully address the exact need of recommendation systems. There is a clear gap between the user’s actual need and the system’s ability to understand it. This gap leads to a poor user experience caused by irrelevant recommendations arising from poorly prioritised factors. This depicts the need for a highly emotion-focused recommendation system, rather than surface-level analysis, in order to make the recommendation system more psychologically grounded.

1.3 Specific Problems

- Current recommendation systems lack the ability to understand or extract the emotional depth of anime content.
- Existing approaches rely heavily on surface-level metadata such as genres, tags, ratings, and user-generated labels.
- Collaborative filtering suffers from cold-start limitations and often produces repetitive or generic recommendations, especially for new users or less popular anime titles.
- Most systems are unable to incorporate or respond to user-defined emotional preferences.
- Conventional methods do not analyse anime dialogues, which are the primary carriers of emotional expression, leading to inaccurate or disconnected recommendation outputs.
- There is no existing recommendation framework that prioritises emotional understanding.

1.4 Aim

This research aims to develop an emotion-aware anime recommendation system based on subtitle analysis using NLP techniques, and then to perform aggregation of each dialogue to obtain an episode-level summary in order to accurately calculate the episode-level emotional profile. The intention is to develop a hybrid approach that recommends the top five anime episodes based on

user-preferred emotional percentages. Additional features expected to be integrated into the hybrid approach include weighted emotional similarity and semantic relevance. The final goal of this hybrid approach is to develop a psychologically grounded recommendation system that will offer a good user experience with more accurate recommendations.

1.5 Research Objectives

The primary contributions of this study are as follows:

1. To perform episode-level emotion classification of anime dialogues, generating an aggregated emotion distribution dataset for all episodes.
2. To compare different modelling approaches for anime recommendation using emotion profiles and semantic embeddings, in order to identify the best-performing approach.
3. To develop and evaluate a hybrid recommendation framework combining weighted emotion similarity and semantic embeddings to enhance recommendation relevance.

2 Literature Review

2.1 Overview of the Anime Industry and Recommendation Need

The rise of Japanese anime from a niche subculture to a worldwide media phenomenon is one of the most significant cross-cultural changes in modern entertainment history. Anime originated in Japan and represents an engaging method of storytelling through an animated approach. The extraordinary ability embedded in anime is the blending of visually attractive and dynamic art styles with sophisticated, emotionally charged stories and engaging character conversations. Viewers align deeply with the story and characters as a result of the engaging nature that anime possesses [8, 9].

A distinct difference between traditional western media and anime lies in a phenomenon called para-social interaction. Japan's culturally rooted sensitivity can be observed in anime content. Through culturally sensitive narratives and the complex journeys of characters—emphasising ethical challenges—a strong relationship forms between characters and viewers, different from traditional Western media. This unintentional relationship leads to an emotionally strong connectivity between two parties, which is what is known as the para-social interaction phenomenon [10]. These fostered intimate relationships create a solid sense of collective fandom and a creative interactive fan base that shares internal emotions and thoughts on digital platforms [8, 11].

Through streaming platforms such as Netflix, geographic and cultural limitations were overcome and previously restricted anime content became accessible [13]. However, this vast accessibility also presented challenges. Users became overwhelmed by the enormous amount of anime available and faced difficulty finding content that matched their emotional inclinations and personal choices [14]. Because of these issues, a recommendation system became prominently necessary [15, 16].

2.2 Anime Recommendation Systems

Modern digital media platforms are now widely used throughout the world. Anime recommendation systems exist as a prominent component in these platforms. Traditional recommendation systems mainly use three types of approaches: content-based filtering, collaborative filtering, and hybrid approaches [31].

2.2.1 Content-Based Filtering

Content-based filtering is used for instances where consistent viewing patterns can be observed. This recommendation approach is based on metadata such as genres, synopses, keywords, and stylistic elements. Despite working well for individuals with regular viewing habits, it struggles with subtle story factors that greatly impact anime appeal. Recent research highlights that relying on simple metadata matching is insufficient for accurate filtering [32], and that semantic features extracted through natural language descriptions can significantly improve the media recommendation domain [33].

2.2.2 Collaborative Filtering

Collaborative filtering identifies patterns by examining user-item engagement and makes suggestions through similarity matrices. It has been advanced into Neural Collaborative Filtering, which aligns with deep learning and latent representation learning [34]. However, cold-start issues and sparse data remain limitations [35], as does the failure to include elements unique to anime—such as emotional tone or complex story development—which are frequently critical to user engagement [36].

2.2.3 Hybrid Approaches

To overcome these drawbacks, hybrid strategies were developed that merge existing approaches [37]. These models integrate item semantics, user behaviour, and contextual data simultaneously [38, 39]. Recent studies have shown that combining embedding-based similarity metrics with user rating patterns greatly increases suggestion relevance for anime [41]. Knowledge graphs and graph neural networks (GNNs) have also been investigated for capturing high-level contextual linkages such as thematic similarity and character archetype overlap [42, 43].

2.3 Emotion Detection in Anime Dialogues

In anime, emotion plays a crucial role because of its tight relationship to character and narrative immersion. The development of precise emotion recognition algorithms for anime dialogues is essential to the development of more personalised recommendation systems.

2.3.1 Transformer Paradigm in Emotion Classification

In initial stages, lexicon-based approaches were used for simple word-to-feeling matching, but due to poor performance when handling sophisticated linguistic phenomena, those models became inadequate [46]. Although traditional models such as SVM, Naive Bayes, and decision trees enhanced categorisation, they had limited capacity for multi-turn conversations [47]. Deep learning architectures using CNNs, LSTMs, and GRUs further improved sequential and semantic pattern extraction [48, 49]. However, all these improvements were unable to capture the emotional depth and accurate structure characteristic of anime dialogue.

The emotion detection field underwent a major transformation with transformer-based models such as BERT, RoBERTa, and DistilBERT. Because of their bidirectional self-attention processes, these models show clear superiority over conventional deep learning models through higher F1 scores and enhanced cross-domain generalisation [50, 51].

2.3.2 Empirical Performance and Domain-Specific Validation

Transformers consistently achieve the best results across fine-grained benchmarks like GoEmotions [52]. Research indicates that transformer topologies can detect emotions from anime subtitles with up

to 91% accuracy, outperforming CNN and LSTM baselines by 12–18% [50]. For Japanese emotion classification on the WRIME corpus, the improved DeBERTa-v3-large model obtained the highest mean F1 score of 0.662, greatly outperforming even strong general models like ChatGPT-4o [55].

The final model selected for emotional level classification of dialogues is the condensed, highly efficient variant `j-hartmann/emotion-english-distilroberta-base`. Key efficiency and scalability considerations form the basis of this decision. Empirical data based on research that introduced the basic DistilBERT architecture confirms that distilled versions are 40% smaller and 60% faster while maintaining 97% of the performance of their larger counterparts [56]. Because anime episodes have thousands of conversation lines, scalability efficiency is critical for high-throughput large-scale processing.

2.4 Hybrid Emotion and Semantic Recommendation Models

2.4.1 Incorporating Emotion Features

The study of emotional data from user behaviour forms the foundation of emotion-aware suggestion. Unlike fundamental information, emotion vectors capture intricate psychological elements like intensity and connection that have a significant impact on the viewing experience. Research confirms that using emotional distributions significantly improves the precision and personalisation of video recommendation systems [58, 59].

2.4.2 Semantic Embeddings for Narrative Understanding

The `all-MiniLM-L6-v2` model will be used in this project to generate semantic representations. When compared to larger BERT or RoBERTa models, MiniLM embeddings maintain good performance while being computationally efficient. Research demonstrates that MiniLM performs exceptionally well on semantic similarity benchmarks, making it ideal for episode descriptors and anime summaries [61, 62].

2.4.3 Machine Learning Models in the Hybrid Pipeline

This framework compares six machine learning models, each chosen for its capacity to handle the structured emotion and semantic information:

AutoML Hybrid Model. AutoML (FLAML, AutoGluon) takes care of the challenging aspects of creating AI models, frequently outperforming manual attempts [63, 64].

LightGBM Hybrid Model. Through leaf-wise gradient boosting combined with histogram optimisation, LightGBM achieves fast training and reliable results on challenging tabular datasets [65].

XGBoost Hybrid Model. XGBoost is one of the best gradient boosting techniques for regression and recommendation problems, capable of handling complex relationships and noisy data [66].

HistGradientBoosting Hybrid Model. HGB is a quick and memory-efficient technique, ideal for handling noisy or sparse data typical in recommendation systems [67].

Mini-TabNet Hybrid Model. TabNet is an artificial intelligence network designed for structured data that employs attention to identify the most crucial information [68, 69].

TabPFN Hybrid Model. TabPFN is a pre-trained model enabling extremely accurate predictions almost instantaneously for small to medium datasets [70, 71].

2.5 Research Gap

Despite the rapid growth of anime-related data and advances in recommender system technologies, several major obstacles remain:

- **Absence of emotion-driven recommendation models:** The majority of anime recommendation systems concentrate on fundamental information such as user preferences, ratings, and genres, rarely making use of the intricate emotional clues contained in the language.
- **Limited hierarchical emotion modelling:** While some studies examine emotions at the dialogue level, there are no multi-level emotional profiles linking feelings across dialogues and entire episodes.
- **Weak integration between semantic and emotional understanding:** Advanced AI models capable of comprehending plot and narrative are rarely paired with in-depth emotional data.
- **Underutilisation of advanced machine learning frameworks:** Techniques like AutoML, TabNet, and TabPFN, perfect for merging many kinds of data, have not yet been incorporated into anime recommendation systems.

3 Methodology

3.1 Introduction

The development of the proposed emotion-aware hybrid anime recommendation engine began with preprocessing anime episode transcripts, followed by the use of simulated user emotion preference profiles to model user interaction for recommendation evaluation. An emotion detection model was then applied to analyse dialogue content, estimate emotional valence, and generate fine-grained emotion vectors. These vectors are aggregated to produce representative emotion profiles at the episode level. The extracted emotional features are subsequently integrated with semantic relevance information and simulated collaborative preference signals to construct hybrid regression-based models that estimate user preference relevance scores. Finally, a hybrid scoring and retrieval pipeline ranks and recommends the most suitable anime episode to users based on these estimated preference scores.

3.2 Data Collection and Preprocessing

The primary dataset consists of more than 1.5 million conversation lines extracted from subtitle files of thousands of anime episodes. The subtitle corpus comprises English-translated anime dialogues, making it suitable for transformer-based emotion detection models trained on English text. To preserve contextual integrity, dialogue lines were carefully extracted during preprocessing while maintaining the episode-to-line mapping. Standard text-cleaning procedures were applied: empty lines were removed, leading and trailing whitespace was trimmed, and punctuation was retained. Due to the scale of the dataset, a batch-processing strategy was employed, where tokenisation and model inference are conducted independently on manageable data chunks. Interim episode-level CSV files were generated and later combined, enabling parallel or incremental execution and ensuring memory-efficient processing.

3.3 Emotion Detection

For each anime episode, emotion probabilities are inferred at the dialogue level, where every dialogue line is assigned a seven-dimensional emotion probability distribution corresponding to predefined emotion categories. The seven emotion classes are: anger, disgust, fear, joy, sadness, surprise, and neutral. Dialogue-level emotion probabilities are aggregated by computing their arithmetic mean across all dialogues within an episode to obtain an overall emotional representation. This aggregation produces a fixed-length, seven-dimensional emotion percentage vector representing the episode's predominant emotional traits and overall emotional composition. These episode-level emotion vectors

serve as the primary emotional feature inputs for the proposed emotion-aware recommendation framework.

3.4 Aggregation and Feature Engineering

Once the seven-dimensional emotion probabilities for each dialogue have been obtained, the data are organised by episode. A representative episode-level emotion profile is generated by calculating the arithmetic mean of each emotion probability across all associated dialogues.

In addition to the six active raw emotion probabilities, three engineered "Anchor Features" are derived from each episode's aggregated emotion profile. These features are designed to provide structurally stable input signals for the downstream regression models, particularly under noisy or imprecise user input conditions:

- **Total Intensity:** Computed as the arithmetic sum of all six active emotion probabilities. This value captures the overall affective energy of an episode and remains relatively stable even when individual emotion scores fluctuate due to subjective reporting variability.
- **Emotion Breadth:** Computed as the count of emotion dimensions whose probability value exceeds a threshold of 0.10. This feature reflects the diversity of emotional expression within an episode, distinguishing emotionally complex episodes from those dominated by a single affect.
- **Positivity Index:** The net difference between positive and negative affect provides an aggregate measure of emotional valence. Because it is derived from the sum of multiple emotion scores rather than any single value, it acts as a stabilisation layer—remaining relatively consistent even when individual components such as Joy or Sadness carry minor noise from user self-reporting.

These three anchor features are appended to the six raw emotion probability values, forming a nine-dimensional input vector for each episode that is passed to the hybrid regression ensemble.

Alongside emotion features, semantic embeddings are computed using the MiniLM model (`all-MiniLM-L6-v2`), generating 384-dimensional embeddings to encode narrative and contextual information derived from episode descriptions. The combined emotional and semantic features form the core input representation for the hybrid recommendation system during the final scoring and retrieval phase.

3.5 Optimisation of Semantic Grounding

This phase introduces a specialised technical layer designed to bridge the gap between raw numerical emotional tensors and human-readable narratives. The optimisation process involves a deterministic transformation of aggregated emotional features into a synthesised "Vibe Column," which serves as a stable semantic fingerprint for every anime episode in the database. Simultaneously, the retrieval architecture is enhanced through an Automatic Semantic Query Generator that translates user-defined emotional slider values into a structured narrative identical in syntax to the database metadata. By implementing this symmetric retrieval framework, the system ensures that cosine similarity calculations are performed between mathematically and linguistically aligned vectors, thereby minimising the semantic gap and improving the contextual relevance of the final recommendations.

The synthesis engine operates through a hierarchical pipeline:

1. **Input and Dimension Isolation:** The process begins by ingesting the six aggregated emotion probabilities (Anger, Disgust, Fear, Joy, Sadness, and Surprise) derived from the DistilRoBERTa transformer model. The "Neutral" class is intentionally excluded to focus the narrative purely on active emotional triggers.

2. **Ordinal Ranking Logic:** The engine performs a priority sort on these six values to identify the Primary Emotional Catalyst (Rank 1), Secondary Influences (Ranks 2–3), and Trace Elements (Ranks 4–6). This hierarchy preserves the emotional depth of the episode beyond a single-label classification.
3. **Linguistic Discretisation Layer:** Continuous numerical values are mapped to qualitative intensity descriptors based on fixed thresholds: *Extreme* (> 75), *High* (50–75), *Moderate* (20–50), and *Subtle* (< 20).
4. **Syntactic Template Assembly:** Descriptors and ranked emotion labels are plugged into a standardised narrative template, ensuring that the `all-MiniLM-L6-v2` model receives information in a consistent, noise-free format.
5. **Vectorisation (Output):** The final narrative is passed to the SentenceTransformer to generate 384-dimensional dense semantic embeddings representing the "emotional fingerprint" used for high-precision matching in the hybrid recommendation pipeline.

3.6 Hybrid Regression Models

To enhance recommendation accuracy beyond direct emotion similarity, six hybrid regression models are trained to predict a quantitative Emotional Affinity Score based on episode attributes: AutoML Hybrid, LightGBM Hybrid, XGBoost Hybrid, HistGradientBoosting Hybrid, Mini-TabNet Hybrid, and TabPFN Hybrid. Each model accepts the six normalised emotion features as primary inputs (Anger, Disgust, Fear, Joy, Sadness, and Surprise). The "Neutral" class is intentionally excluded from model inputs.

Training is conducted using an Emotional Affinity Score—a theoretically grounded composite measure derived from weighted emotion probabilities, where weights reflect the relative contribution of each emotion dimension to viewer engagement: joy (0.7), sadness (0.7), anger (0.6), fear (0.6), surprise (0.6), and disgust (0.5). Gaussian jitter ($\sigma = 2.0$) is applied to simulate the natural subjectivity of human emotional self-reporting. An 80/20 train–test split and standard regression loss functions are employed. Model performance is evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score.

Critically, all six models are subjected to Gaussian noise injection at incremental levels from 0% to 10%, simulating the natural variability of user-specified emotion slider values. A Noise Resilience Score (NRS), computed as the mean R^2 across all noise levels, determines final model selection. The model achieving the highest NRS is selected as the final regression backbone, even if it does not achieve the highest clean-data R^2.

3.7 Hybrid Scoring and Retrieval

A composite hybrid scoring mechanism generates final recommendations by integrating multiple preference signals. The system automatically translates user numerical emotion slider inputs into a structured narrative query that mirrors the exact syntactic format of the "Vibe Column" synthesised during preprocessing. Cosine similarity is computed between the user-specified emotion vector and the episode's aggregated emotion profile. Simultaneously, symmetric semantic similarity is calculated between the pre-computed episode vibe embedding and the automatically generated user mood embedding using `all-MiniLM-L6-v2`. Additionally, a learned preference estimate is obtained from the trained hybrid regression model. The final recommendation score is computed as a weighted combination of these components. Episodes are then ranked, and the top five results are retrieved as the final set of multifaceted recommendations.

3.8 Technical Implementation Stack

To ensure reproducibility and technical clarity, Table 1 summarises the primary libraries and frameworks utilised throughout the research methodology.

Table 1: Technical Stack

Methodology Phase	Primary Technology	Functions
Data Preprocessing	Python (Pandas/NumPy)	Batch processing and CSV aggregation
Emotion Detection	DistilRoBERTa (Transformers)	7-dimensional dialogue sentiment inference
Semantic Encoding	all-MiniLM-L6-v2	384D Vibe Embedding generation
Regression Ensemble	AutoML, LGBM, XGB, HGB, TabNet, TabPFN	Multi-model benchmarking suite
Similarity Metrics	Scikit-learn (Cosine Similarity)	Vector distance calculations for retrieval
Stress Testing	NumPy (Gaussian Noise)	Robustness evaluation via noise injection

3.9 Evaluation and Reproducibility

A comprehensive two-pronged evaluation strategy is adopted. The quantitative component utilises offline regression-based metrics, specifically MSE and R^2, to provide a mathematical measure of prediction accuracy. The methodology incorporates a detailed comparative analysis between training and testing performance to validate the efficacy of the stochastic noise injection and structural regularisation strategies employed during ensemble training. Furthermore, the system undergoes qualitative human evaluation to gauge recommendation satisfaction and the degree of emotional congruence between the user's articulated state and the suggested content. Random seeds are strictly fixed across all computational libraries, and all intermediate CSV outputs are archived with comprehensive provenance metadata.

4 Results and Evaluation

4.1 Impact of the Anchor Feature

A central contribution of this research is the longitudinal comparison between a baseline 6D vector (comprising the raw emotions of Anger, Disgust, Fear, Joy, Sadness, and Surprise) and an engineered 9D vector incorporating three "Anchor Features."

In initial testing using only raw emotional probabilities, attention-based models established the performance ceiling on clean data but demonstrated extreme vulnerability to input variance. In the 6D environment, negative R^2 values at 10% noise indicated a critical failure for real-world deployment, as models performed worse than simply predicting the dataset mean. The 6D performance baseline is summarised in Table 2.

Table 2: 6D Performance Baseline

Model	**CV Mean R^2**	**Test R^2**	**MAE**	**NRS**	**R^2 @ 10% Noise**
TabPFN	0.9738	0.9758	1.5843	0.5301	-0.2113
TabNet	0.9622	0.9706	1.7104	0.5640	-0.0804
XGBoost	0.9245	0.9314	2.5473	0.6766	0.2908
HGB	0.9169	0.9235	2.6662	0.6684	0.2798
LightGBM	0.9180	0.9235	2.6635	0.6679	0.2778

The introduction of the 9D input vector produced a decisive improvement in operational stability across all architectures, particularly in terms of the Noise Resilience Score (NRS). By augmenting the raw 6D emotion probabilities with Total Intensity, Emotion Breadth, and the Positivity Index, the system transitioned from a state of high vulnerability to one of robust industrial readiness. Results are presented in Table 3.

Table 3: 9D Performance Matrix After Integration of Anchor Features

Model	**Train R^2**	**Test R^2**	**Gen.Gap**	**MAE**	**MSE**	**NRS**	**R^2 @ 10%**
TabPFN	0.9758	0.9758	-0.0000	1.5886	4.0290	0.8735	0.7227
TabNet	0.9682	0.9704	-0.0022	1.7168	4.9387	0.8979	0.7792
HGB	0.9725	0.9698	0.0027	1.7145	5.0355	0.8944	0.7626
LightGBM	0.9725	0.9701	0.0025	1.7002	4.9877	0.8947	0.7631
XGBoost	0.9756	0.9699	0.0057	1.7289	5.0136	0.8980	0.7999

The comparative data highlights a significant "Robustness Recovery." The most dramatic improvements were observed in the attention-based models: TabPFN showed a 65% relative improvement and TabNet improved by 59%. This recovery is attributed to the "Stabilisation Layer" provided by the anchor features. While raw emotion dimensions like "Joy" or "Fear" might fluctuate significantly due to subjective user reporting, Total Intensity and Positivity Index remain mathematically stable, allowing models to maintain focus on underlying emotional energy rather than surface-level noise.

4.2 Baseline Performance in the 9D System

On clean data, TabPFN established the performance ceiling for the 9D system, maintaining a Test R^2 of 0.9758 with an effectively zero generalisation gap. The gradient boosting ensemble clustered

tightly: XGBoost (0.9699), LightGBM (0.9701), and HGB (0.9698), confirming that the 9D vector provides a high-performance floor across all architectures. Notably, the AutoML (FLAML) framework independently identified XGBoost as the optimal internal learner for this affective dataset across 141 search iterations.

4.3 Robustness Profiling and Attention Collapse

The stress test revealed a decisive architectural divide. While attention-based architectures previously exhibited "Attention Collapse" beyond the 4% noise threshold in the 6D baseline, the 9D configuration successfully reversed this failure across the ensemble. At the 10% noise ceiling, both attention-based models declined into negative R^2 territory in the 6D system, indicating complete failure for real-world deployment. In contrast, tree-based models such as XGBoost and HGB maintained a positive R^2 profile across the entire noise spectrum. By expanding the feature space with structurally stable aggregates, all five architectures maintained positive R^2 scores even at maximum noise, with XGBoost achieving the strongest adversarial performance at $R^2 = 0.7999$ under 10% noise.

4.4 Feature Engineering Ablation of Anchor Features

The superior stability of the 9D system is directly attributed to the three engineered Anchor Features. Total Intensity captures aggregate affective energy and remained a constant, dominant influence on the recommendation score even as individual emotional probabilities fluctuated. The Positivity Index provides a structural anchor by calculating the net difference between positive and negative affective states, effectively smoothing out individual slider errors. While raw emotional inputs exhibit high sensitivity to input jitter, these anchor features maintain constant influence, effectively filtering out human reporting noise while preserving the dominant emotional signal.

4.5 Hybrid Synergy: Semantic Vibe and Retrieval Validation

The final evaluation phase assessed the synchronisation between the numerical tabular models and the Sentence-Transformer (`all-MiniLM-L6-v2`) semantic layer. This hybrid synergy successfully addresses the "Tabular Cold Start" problem, where multiple episodes might yield near-identical regression scores. The system converts raw emotional tensors into a qualitative "vibe description" using intensity descriptors (Extreme, High, Moderate, Subtle), bridging the gap between mathematical vectors and human language. The Semantic Layer acts as a deterministic tie-breaker: when the tabular model assigns identical preference scores, the system selects the recommendation with the highest narrative cosine similarity. Approximately 80% inter-model agreement was observed across all six architectures for identical user inputs, confirming stable and emotionally coherent recommendation behaviour.

5 Discussion

5.1 Summary of Findings

This research successfully developed an emotion-aware hybrid anime recommendation system that combines subtitle-based emotional analysis with semantic embeddings to produce psychologically grounded episode-level recommendations. The `j-hartmann/emotion-english-distilroberta-base` transformer model was applied to over 1.5 million dialogue lines to generate seven-dimensional episode-level emotion profiles, directly addressing the absence of hierarchical emotion modelling in existing anime recommendation pipelines.

A six-model benchmarking study revealed that model selection criterion matters as much as architecture. While TabPFN achieved the highest clean-data accuracy, it suffered catastrophic degradation under noise in the 6D baseline. The introduction of three engineered Anchor Features expanded the input space from 6D to 9D and produced a decisive improvement in noise resilience across all architectures, with NRS gains ranging from 33% for gradient boosting models to 65% for TabPFN. XGBoost emerged as the optimal backbone with the highest NRS of 0.8980, a finding independently validated when the AutoML framework selected it as its best internal learner. The hybrid scoring pipeline combining the 9D XGBoost regression model with the `all-MiniLM-L6-v2` Sentence-Transformer semantic layer demonstrated stable and emotionally coherent recommendation behaviour.

5.2 Limitations of the Study

While the system demonstrated strong empirical performance, several limitations were identified:

- **Training Data Constraints:** In the absence of labelled user data, models were trained using simulated preference scores derived from weighted episode emotion probabilities; consequently, the system approximates a mathematical formula rather than behavioural user preferences.
- **Translation and Linguistic Nuance:** Reliance on English-translated subtitles introduces a translation layer that may fail to capture culturally specific Japanese emotional concepts or subtle affective nuances present in the original source material.
- **Evaluation Scope:** The current validation framework is entirely offline and quantitative; the absence of live user evaluation means that subjective satisfaction and perceived emotional congruence have not been verified in a real-world environment.
- **Granularity and Adaptability:** The system operates strictly at the episode level and utilises static emotion profiles, preventing it from generating series-level recommendations or adapting to shifting user emotional trajectories across multiple viewing sessions.

5.3 Recommendations for Future Work

The most important future direction is the replacement of the simulated Emotional Affinity Score with genuine user preference data, potentially collected through a deployment study on anime streaming communities. Future work should also investigate emotion detection directly on Japanese-language subtitle text, eliminating the translation layer. Models such as the DeBERTa-v3-large fine-tuned on the WRIME corpus, which achieved a mean F1 score of 0.662 on Japanese emotion classification tasks, represent a strong starting point [55]. The current episode-level granularity could be extended upward to produce series-level and story-arc-level affective profiles. Finally, future work could incorporate additional modalities including audio features (voice tone, music, sound effects) and visual features (scene colour palette, facial expressions, animation style).

6 Conclusions

This research successfully addressed a significant gap in affective computing by developing a psychologically grounded, noise-resilient anime recommendation framework that achieved all three primary research objectives:

a. The `j-hartmann/emotion-english-distilroberta-base` transformer was deployed at scale to perform episode-level emotion classification across over 1.5 million anime dialogue lines, generating a seven-dimensional emotion profile dataset that addresses the absence of hierarchical emotion modelling in existing pipelines.

b. A six-model benchmarking study identified XGBoost as the optimal regression backbone, with an NRS of 0.8980—a finding independently validated by the AutoML framework across 141 search iterations—demonstrating that noise robustness is a more meaningful selection criterion than peak clean-data accuracy for affective computing systems.

c. The transition from a 6D to a 9D input space, facilitated by engineered Anchor Features (Total Intensity, Emotion Breadth, and Positivity Index), produced NRS improvements of 33% to 65% across all architectures, effectively transforming previously vulnerable models into systems ready for real-world deployment.

Although limited by the use of simulated preference data and an English-language dependency, this foundational architecture provides a technically robust platform for future multimodal affective extensions in the domain of personalised entertainment.

References

[1] Precedence Research. (2024). Anime market size, share, growth, and forecast to 2034. Retrieved from https://www.precedenceresearch.com/anime-market

[2] GamesRadar. (2025). Netflix says that over 50% of subscribers watch anime. Retrieved from https://www.gamesradar.com/entertainment/anime-shows/netflix-says-that-over-50-percent-of-subscribers-watch-anime-and-its-all-thanks-t

[3] Sarhan, A. M., Ayman, H., Wagdi, M., Ali, B., Adel, A., & Osama, R. (2024). Integrating machine learning and sentiment analysis in movie recommendation systems. *Journal of Electrical Systems and Information Technology*, 11(1), 53. https://doi.org/10.1186/s43067-024-00177-7

[4] Vaswani, A., et al. (2023). Attention is all you need. *arXiv:1706.03762*. https://doi.org/10.48550/arXiv.1706.03762

[5] Kane, A., Patankar, S., Khose, S., & Kirtane, N. (2022). Transformer based ensemble for emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis* (pp. 250–254). https://doi.org/10.18653/v1/2022.wassa-1.25

[6] Pereira, P., Moniz, H., & Carvalho, J. P. (2024). Deep emotion recognition in textual conversations: A survey. *arXiv:2211.09172*. https://doi.org/10.48550/arXiv.2211.09172

[7] Hara, Y., & Itou, K. (2010). Classification of emotional speech in anime films by using automatic temporal segmentation.

[8] Mahamood, A. F., Haris, M. I. I., Yakob, T. K. T., Ramli, A. J., & Ahmad, Z. (2024). Exploring the impact of social media on anime fandom. Vol. 10, No. 28.

[9] Hajek, A., & König, H.-H. (2024). Interest in anime and manga: Relationship with (mental) health, social disconnectedness, social joy and subjective well-being. *Journal of Public Health*. https://doi.org/10.1007/s10389-024-02341-9

[10] Ramu, V., & Tanwar, K. Impact of para-social interaction with anime characters on self-esteem and subjective happiness in young adults.

[11] YouWei Trade. (2025). What does psychology say about people who watch anime? Retrieved from https://youweitrade.com/blogs/blog/what-does-psychology-say-about-people-who-watch-anime-unpacking-fandom-identity-ar

[12] ArticleAlley. (2025). The anime effect: Exploring its pervasive role in international pop culture. Retrieved from https://www.articlealley.com/news/the-anime-effect-exploring-its-pervasive-role-in-international-pop-culture-23082

[13] Petit, A., et al. (2022). Anime streaming platform wars: A platform lab report. Unpublished. https://doi.org/10.13140/RG.2.2.34667.67368

[14] Wang, D., & Zhao, X. (2022). Affective video recommender systems: A survey. *Frontiers in Neuroscience*, 16, 984404. https://doi.org/10.3389/fnins.2022.984404

[15] Krishna, E. S. P., et al. (2025). Enhancing e-commerce recommendations with sentiment analysis using MLA-EDTCNet and collaborative filtering. *Scientific Reports*, 15(1), 6739. https://doi.org/10.1038/s41598-025-91275-7

[16] Rajput, H. (2025). Emotion based music and video recommendation system. *International Journal of Research in Applied Science and Engineering Technology*, 13(4), 4745–4751. https://doi.org/10.22214/ijraset.2025.68246

[17] Sarkar, S. (2025). Development of a hybrid recommendation system using collaborative filtering and content-based filtering techniques. Unpublished. https://doi.org/10.13140/RG.2.2.32688.47369

[18] Zhang, K., et al. (2021). SIFN: A sentiment-aware interactive fusion network for review-based item recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (pp. 3627–3631). https://doi.org/10.1145/3459637.3482181

[19] Harputlu Yamak, Y., & Işık, Y. (2024). Anime watching: Is a new kind of addiction? *Middle East Current Psychiatry*, 31(1), 73. https://doi.org/10.1186/s43045-024-00463-0

[20] Zhang, X., et al. (2024). Towards empathetic conversational recommender systems. In *18th ACM Conference on Recommender Systems* (pp. 84–93). https://doi.org/10.1145/3640457.3688133

[21] Kim, S., Kim, H., Lee, J., Jeon, Y., & Lee, G. G. MIRROR: Multimodal cognitive reframing therapy for rolling with resistance.

[22] Wu, C., et al. (2025). Multimodal emotion recognition in conversations: A survey. *arXiv:2505.20511*. https://doi.org/10.48550/arXiv.2505.20511

[23] Zhang, S., et al. (2024). EEG-SVRec: An EEG dataset with user multidimensional affective engagement labels in short video recommendation. *arXiv:2404.01008*. https://doi.org/10.48550/arXiv.2404.01008

[24] Tkalčič, M., Burnik, U., Odić, A., Košir, A., & Tašič, J. (2013). Emotion-aware recommender systems – a framework and a case study. In *ICT Innovations 2012*, Advances in Intelligent Systems and Computing, Vol. 207 (pp. 141–150). https://doi.org/10.1007/978-3-642-37169-1_14

[25] Brodbeck, C., Hannagan, T., & Magnuson, J. S. (2024). Recurrent neural networks as neurocomputational models of human speech recognition. https://doi.org/10.1101/2024.02.20.580731

[26] Lind, M. N., Byrne, M. L., Wicks, G., Smidt, A. M., & Allen, N. B. (2018). The Effortless Assessment of Risk States (EARS) tool. *JMIR Mental Health*, 5(3), e10334. https://doi.org/10.2196/10334

[27] Jing, E., et al. (2024). Emotion-aware personalized music recommendation with a heterogeneity-aware deep Bayesian network. *arXiv:2406.14090*. https://doi.org/10.48550/arXiv.2406.14090

[28] Babu, T., Nair, R. R., & A, G. (2023). Emotion-aware music recommendation system: Enhancing user experience through real-time emotional context. *arXiv:2311.10796*. https://doi.org/10.48550/arXiv.2311.10796

[29] Kanchan, K. G., Mulla, A. Y., & Ramesh, A. S. (2024). AI for emotion based recommendation systems. Vol. 11, No. 1.

[30] Abdul, A., Chen, J., Liao, H.-Y., & Chang, S.-H. (2018). An emotion-aware personalized music recommendation system using a convolutional neural networks approach. *Applied Sciences*, 8(7), 1103. https://doi.org/10.3390/app8071103

[31] Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender Systems Handbook* (pp. 1–35). Springer. https://doi.org/10.1007/978-0-387-85820-3_1

[32] Xu, W., Xie, Q., Yang, S., Cao, J., & Pang, S. (2024). Enhancing content-based recommendation via large language model. *arXiv:2404.00236*. https://doi.org/10.48550/arXiv.2404.00236

[33] Zhao, Z., et al. (2024). Recommender systems in the era of large language models. *IEEE Transactions on Knowledge and Data Engineering*, 36(11), 6889–6907. https://doi.org/10.1109/TKDE.2024.3392335

[34] Rendle, S., Krichene, W., Zhang, L., & Anderson, J. (2020). Neural collaborative filtering vs. matrix factorization revisited. *arXiv:2005.09683*. https://doi.org/10.48550/arXiv.2005.09683

[35] Zhang, S., Yao, L., Sun, A., & Tay, Y. (2020). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1), 1–38. https://doi.org/10.1145/3285029

[36] Koren, Y. (2009). Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD Conference* (pp. 447–456). https://doi.org/10.1145/1557019.1557072

[37] Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis*, 21(6), 1487–1524. https://doi.org/10.3233/IDA-163209

[38] Tian, Y., & Fu, S. (2020). A descriptive framework for the field of deep learning applications in medical images. *Knowledge-Based Systems*, 210, 106445. https://doi.org/10.1016/j.knosys.2020.106445

[39] Li, H., Lin, J., Wang, T., Zhang, L., & Wang, P. (2022). A personalized short video recommendation method based on multimodal feature fusion. *In Review*. https://doi.org/10.21203/rs.3.rs-2033641/v1

[40] Kvifte, T. Video recommendations based on visual features extracted with deep learning.

[41] Wu, Y., et al. (2025). Anime generation through diffusion and language models. *Computer Modeling in Engineering & Sciences*, 144(3), 2709–2778. https://doi.org/10.32604/cmes.2025.066647

[42] van den Berg, R., Kipf, T. N., & Welling, M. (2017). Graph convolutional matrix completion. *arXiv:1706.02263*. https://doi.org/10.48550/arXiv.1706.02263

[43] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. (2020). LightGCN: Simplifying and powering graph convolution network for recommendation. In *43rd International ACM SIGIR Conference* (pp. 639–648). https://doi.org/10.1145/3397271.3401063

[44] Khan, H. U., Naz, A., Alarfaj, F. K., & Almusallam, N. (2025). A transformer-based architecture for collaborative filtering modeling. *Scientific Reports*, 15(1), 24503. https://doi.org/10.1038/s41598-025-08931-1

[45] Liu, H., Wei, Y., Song, X., Guan, W., Li, Y.-F., & Nie, L. (2024). MMGRec: Multimodal generative recommendation with transformer model. *arXiv:2404.16555*. https://doi.org/10.48550/arXiv.2404.16555

[46] Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1), 81. https://doi.org/10.1007/s13278-021-00776-6

[47] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

[48] Singh, V., Tripathi, N. K., & Jain, R. (2025). Multi-facial emotion recognition using fusion CNN. Vol. 1, No. 1.

[49] Mamani-Coaquira, Y., & Villanueva, E. (2024). A review on text sentiment analysis with machine learning and deep learning techniques. *IEEE Access*, 12, 193115–193130. https://doi.org/10.1109/ACCESS.2024.3513321

[50] Rezapour, M. Emotion detection with transformers: A comparative study.

[51] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

[52] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv:2005.00547*. https://doi.org/10.48550/arXiv.2005.00547

[53] Hasan, K., Saquer, J., & Ghosh, M. (2025). Advancing mental disorder detection: A comparative evaluation of transformer and LSTM architectures. *arXiv:2507.19511*. https://doi.org/10.48550/arXiv.2507.19511

[54] Kim, S., & Lee, S.-P. (2023). A BiLSTM-Transformer and 2D CNN architecture for emotion recognition from speech. *Electronics*, 12(19), 4034. https://doi.org/10.3390/electronics12194034

[55] Takenaka, Y. (2025). Performance evaluation of emotion classification in Japanese using RoBERTa and DeBERTa. *arXiv:2505.00013*. https://doi.org/10.48550/arXiv.2505.00013

[56] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv:1910.01108*. https://doi.org/10.48550/arXiv.1910.01108

[57] Zhang, K., et al. (2025). Robust recommender system: A survey and future directions. *ACM Computing Surveys*, 58(1), 1–38. https://doi.org/10.1145/3757057

[58] Hasan, T., & Bunescu, R. (2025). A survey of affective recommender systems. *arXiv:2508.20289*. https://doi.org/10.48550/arXiv.2508.20289

[59] Pichappan, P. (2025). A review of the emotion-induced music recommendation systems. *Journal of Digital Information Management*, 23(2), 112–133. https://doi.org/10.6025/jdim/2025/23/2/112-133

[60] Hartmann, J. (2021). `j-hartmann/emotion-english-distilroberta-base`. Hugging Face. Retrieved from https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

[61] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv:2002.10957*. https://doi.org/10.48550/arXiv.2002.10957

[62] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv:1908.10084*. https://doi.org/10.48550/arXiv.1908.10084

[63] Zheng, R., Qu, L., Cui, B., Shi, Y., & Yin, H. (2023). AutoML for deep recommender systems: A survey. *arXiv:2203.13922*. https://doi.org/10.48550/arXiv.2203.13922

[64] Erickson, N., et al. (2020). AutoGluon-Tabular: Robust and accurate AutoML for structured data. *arXiv:2003.06505*. https://doi.org/10.48550/arXiv.2003.06505

[65] Ke, G., et al. LightGBM: A highly efficient gradient boosting decision tree.

[66] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

[67] Ramjee, S., & Wornow, M. Histogram-based gradient boosting trees for efficient graph learning with Wasserstein embeddings.

[68] Arik, S. O., & Pfister, T. (2020). TabNet: Attentive interpretable tabular learning. *arXiv:1908.07442*. https://doi.org/10.48550/arXiv.1908.07442

[69] Yuan, Y., et al. (2024). ContextGNN: Beyond two-tower recommendation systems. *arXiv:2411.19513*. https://doi.org/10.48550/arXiv.2411.19513

[70] Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2023). TabPFN: A transformer that solves small tabular classification problems in a second. *arXiv:2207.01848*. https://doi.org/10.48550/arXiv.2207.01848

[71] Zeng, Y., Dinh, T., Kang, W., & Mueller, A. C. (2025). TabFlex: Scaling tabular learning to millions with linear attention. *arXiv:2506.05584*. https://doi.org/10.48550/arXiv.2506.05584