TECHNIUM
SOCIAL SCIENCES JOURNAL

Vol. 81/2026

A New Decade for Social Changes

PLUS COMMUNICATION
International Communication & PR

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

White-Boxing and Personalization of Music Recommendation Models through Attention Visualization and Retraining

Tasuku Okada[1*], **Hiromitsu Shimakawa**[2], **Fumiko Harada**[3]

[1, 2, 3] Ritsumeikan University

*Corresponding author: Tasuku Okada, is0549vx@ed.ritsumei.ac.jp

Abstract. The widespread adoption of music streaming services allows users to access a vast number of songs with ease. However, identifying music that matches a user's situational context or emotional state remains challenging. Recent text-based recommendation approaches using models such as BERT have made significant progress. However, when recommendations fail, users often cannot identify the reasons for failure, resulting in a recognition gap between users and the model. To address the issue, the study proposes an explainable and user-adaptive music recommendation method. The proposed method enables users to understand recommendation rationales and refine the model through feedback. First, the model is pre-trained using Triplet Loss on YouTube comments to capture music-related nuances. It then performs classification-based learning that takes user comments as input and outputs song recommendations. The method visualizes self-attention weights in BERT as heatmaps. The visualizations highlight the words the model focuses on during the recommendation process. Based on the visualizations, users can specify words to strengthen or weaken. The model is retrained using feedback-driven Triplets, enabling modification of attention distributions and recommendation results. Experiments have verified changes in attention distribution due to different learning methods, differences in recommendation results for similar queries, and human evaluation. Their results show that attention modification for user-specified words succeeds in approximately 60–70% of cases. Under conditions for practical usages, the success rate increases to around 80%. Post-retraining recommendations achieve high user satisfaction for top-ranked results. The highest-ranked recommendation attains an average satisfaction score of 4.15. The proposed method allows users to specify model errors, which directs the model toward a suitable one for their preferences.

Keywords. Attention, Triplet Loss, Retraining, Music Recommendation, BERT, Personalization

1. INTRODUCTION

In recent years, music streaming services such as Spotify[1] and Apple Music[2] have enabled users to access an enormous number of songs with ease. However, the increase in available songs has made it more difficult for users to reach music that matches their personal preferences.

To improve user satisfaction, various methods have been proposed to provide high-quality music recommendations. Traditional music recommendation systems have widely employed collaborative filtering[3], content-based approaches[4], and hybrid methods combining both strategies[5]. However, collaborative filtering relies heavily on historical data and shared preferences, which makes it vulnerable to the cold-start problem for new users and new songs. While content-based recommendation can address song features such as lyrics, acoustics, artist

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

information, and vocal quality, it is difficult to reflect the atmosphere and nuances that users request in recommendations.

Recently, some studies have utilized language models such as BERT for text-based music recommendation to handle ambiguous requests. However, the decision-making processes of deep learning models tend to remain black-boxed[6][7]. Consequently, when recommendations differ from user expectations, users often cannot understand why a song is recommended or which aspects deviated from their intent. The issue manifests not only as insufficient recommendation accuracy but also as a recognition gap between user intent and the model's interpretation[8].

A recognition gap refers to the discrepancy between the aspects emphasized in a user's query and the elements that the model actually focuses on. For example, when a user requests "summer songs for high school students," the model may focus on "summer" while insufficiently considering the user's emphasis on "high school students." As a result, the system may recommend a well-known but outdated summer song, which fails to satisfy the user. Note that conventional methods do not visualize model decision processes. Due to this, the users cannot understand why recommendations fail, which prevents them from conveying which aspects need correction.

The combination of black-box behaviour and recognition gaps reduces user trust and satisfaction, consequently discouraging continued use of recommendation systems. From these observations, existing text-based music recommendation methods still leave serious drawbacks in resolving recognition gaps and improving the interpretability of recommendation rationales. Music recommendation systems must not only be accurate but also explain their reasons. It should help users recognize gaps to give a way to correct them when necessary.

The study proposes a music recommendation model that explains the recommendation reasons. It enables the model to modify aspects to value in the recommendation based on user feedback. For explainability, the method visualizes BERT self-attention to intuitively show which words the model focuses on when generating recommendations. For modifiability, users specify words to reinforce or weaken. The method fine-tunes the model to reduce Triplet Loss, which aligns attention distributions to accommodate recommendations with user intent. Furthermore, to help the model capture music-specific phrasing and nuanced expressions, the study introduces pre-training using Triplet Loss for expression learning. It corresponds to nuance learning, adjusting the semantic representations underlying recommendations.

The paper conducts three experiments to validate the effectiveness of the proposed method. First, the paper compares the attention patterns of BERT without fine-tuning, BERT with only classification learning, and BERT with pre-trained Triplet Loss learning, to analyze the impact of nuance learning and classification learning on attention distribution. Second, the paper verifies whether attention and recommendation results appropriately adapt to queries that are similar in content but differ in nuance, to confirm the proposed method's ability to capture subtle differences in user requests. Third, the paper conducts human evaluation experiments to demonstrate that feedback-driven retraining improves the recommendation experience.

The results show that the proposed method enables users to identify recognition gaps through the visualization of recommendation rationales. In addition, Triplet Loss-based retraining allows attention to be modified toward intended directions with a certain success rate. In particular, attention modification has achieved success rates of 60–70%, rising to about 80% when excluding overly extreme adjustments. Moreover, post-retraining recommendations have achieved high satisfaction scores, with top-ranked items averaging 4.15 points, which confirms improvement of user satisfaction.

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

With the method, users can lead the model to recommend songs that best suit their personal preferences from a vast selection of songs. From then on, users can enjoy stress-free recommendations.

2. Technologies to Recommend Music

2.1. Attention-Based Large Language Models and Word Embeddings

In the field of natural language processing, distributed representations are widely used to represent words and sentences as numerical vectors and to model their semantic relationships.

Traditional methods such as Bag-of-Words and TF-IDF rely on word frequency, making it difficult to adequately capture semantic similarity between words and contextual dependencies. Subsequently, methods such as Word2Vec enabled vector representations in which semantically similar words are positioned close to each other. However, they remained unable to handle words whose meanings vary depending on context.

To address this limitation, Transformer-based large language models (LLMs), such as BERT[9] and GPT[10], were proposed. LLMs are pre-trained on large-scale text corpora and are capable of generating context-aware word representations. In BERT, even the same word can produce different vector representations depending on its context within a sentence, enabling flexible semantic understanding of ambiguous expressions and emotionally nuanced sentences.

The core component of the Transformer architecture is the attention mechanism[11]. Attention computes weights that indicate how strongly each word in a sentence should attend to other words, and updates word representations based on these weights. This mechanism enables the model to dynamically generate representations that emphasize important words while considering the overall semantic structure of the sentence. Furthermore, the self-attention explicitly models dependencies between words, making it effective for handling long sentences and complex contexts.

In LLMs, word vectorization is achieved by stacking multiple attention layers. The final layer then produces embeddings that reflect the overall meaning of a sentence and the relationships among its components. In addition, attention weights can be obtained as numerical values, allowing visualization of which words the model focuses on when making decisions[12]. This property mitigates the black-box nature of deep learning models and contributes to improved explainability.

2.2. Triplet Loss Learning

Triplet Loss learning is a distance-based metric learning method that is widely used to learn embedding representations in fields such as image recognition and natural language processing[13][14]. This method employs triplets consisting of an anchor that is reference data, a positive sample that is semantically similar to the anchor, and a negative sample that is semantically dissimilar. The model is trained so that the distances among these samples satisfy predefined constraints. Specifically, the model is trained under the constraint that the distance between the anchor and the positive is smaller than that between the anchor and the negative. This training objective constructs an embedding space in which semantically similar data are positioned closer together.

In Triplet Loss learning, the selection of negative samples has a significant impact on learning performance. When only easy negatives that are clearly unrelated to the Anchor are used, the distance constraint is easily satisfied, causing training to converge prematurely. As a result, the learned embeddings often fail to distinguish subtle semantic differences. In contrast, hard

negatives share some vocabulary or topics with the Anchor but differ in overall semantic intent. Using such samples forces the model to identify finer semantic distinctions, leading to improved representation accuracy[15].

2.3. Dataset for Implementing a Music Recommendation Model

In music recommendation tasks, the choice of dataset has a significant impact on recommendation accuracy and the scope of analysis. In general, content-based music recommendation ideally integrates diverse information sources, including lyrics, acoustic features (e.g., mel-spectrograms, tempo, and rhythm), artist information, vocal timbre, and singing style. By combining these information sources, recommendations can capture the multifaceted characteristics of songs. However, collecting, preprocessing, and modeling all of these information sources individually is challenging within the limited time and resources.

YouTube comment data [16] has possibilities to solve the problem of multiple sources. YouTube comments contain listeners' impressions and evaluations expressed in natural language, often including references to lyrics, acoustic impressions, vocal quality, and artist identity. For example, comments may describe that "the lyrics really resonate," "the voice sounds gentle," "the melody is bright," or that a song "has a summery vibe."

In this respect, YouTube comments can be regarded as textual data that holistically reflects diverse aspects of a song. It provides a comprehensive view of the multiple sources rather than treating them individually.

3. Proposed Method

3.1. Hypothesis

The study proposes an explainable and modifiable system that recommends music based on user query sentences. The study assumes the attention mechanism has enough interpretability and expressive power for sentences in natural language. The study leverages it not only to explain recommendation rationales but also to support subsequent representation adjustments.

The study aims not to precisely extract objective acoustic features of songs, but to model users' subjective impressions, identifying the elements they emphasize when perceiving music. YouTube comment data, which directly incorporates listeners' linguistic expressions, is highly compatible with the aim of this research. The study uses YouTube comment data as the primary information source for implementing the music recommendation model. The approach enables the exploration of recommendation methods that consider the multifaceted nuances of songs, even under limited research resources.

The study applies the reduction of Triplet Loss to make the embedded space of LLM adapted to better align with the music recommendation task and user intent. In the study, song-related query sentences provided by users are used as anchors. To construct Triplets, YouTube comments semantically similar to the anchors are selected as Positive samples, while comments that share partial expressions or topics but differ in overall intent are selected as hard negatives, together with easy negatives. Through representation learning with Triplet Loss, clarifying the semantic relationships between queries and comments is expected to induce changes in the attention distribution.

Figure 3.1 illustrates an overview of the proposed method. For explaining recommendation rationales, the proposed method utilizes the attention mechanism introduced in Section 2.1. For model modification, the method employs Triplet Loss learning described in Section 2.2. The details of each component in the diagram are explained sequentially below.

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

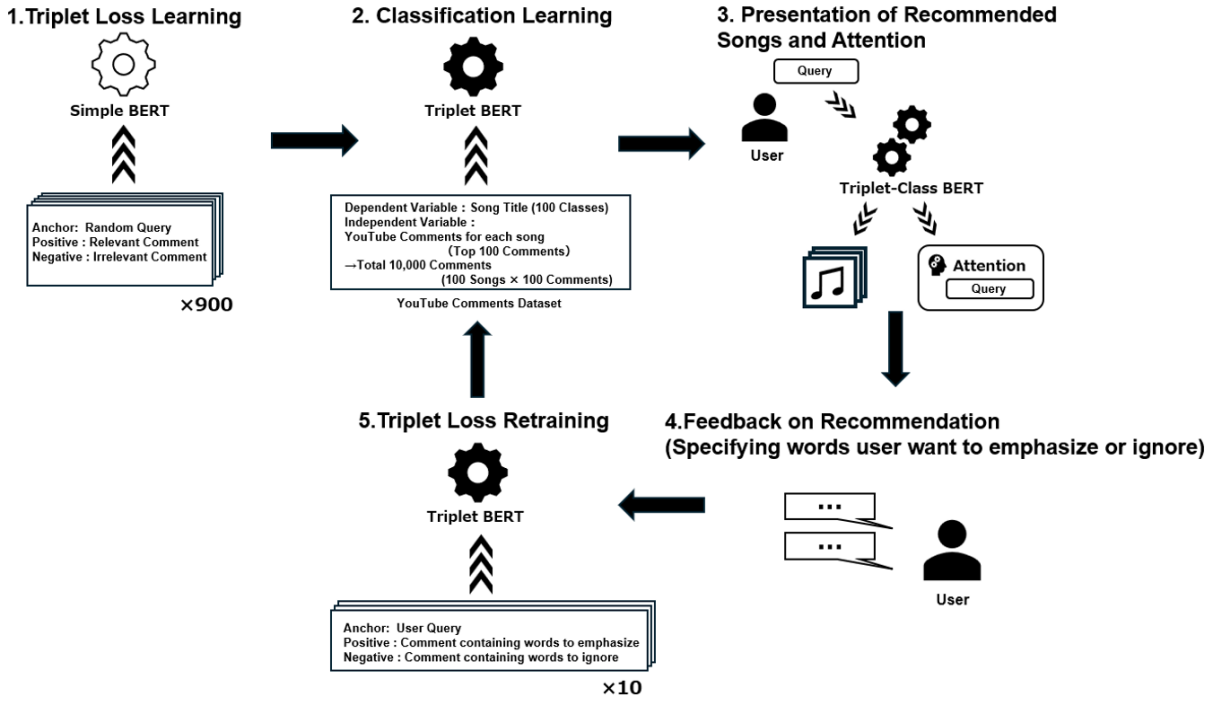

Figure 3.1: Overview of the proposal method

3.2. Attention Mechanism Reflecting Musical Nuances

In recommendations based on content represented in natural language, performance strongly depends on how accurately the model understands user expressions. Each domain contains expressions that involve domain-specific phrasing, meanings, and nuances. However, a general-purpose LLMs designed to process broad and general knowledge often struggle to adequately capture such domain-specific nuances. Accordingly, the study causes an LLM to acquire semantic representations specialized for the music domain. The method preliminarily trains a general BERT model, reducing Triplet Loss, to construct an attention mechanism that reflects musical nuances. Through the training, the model more sensitively captures the semantic structure of input sentences. A corresponding changes emerge in its attention mechanism.

Regarding the training data, the study prepares 100 randomly generated query sentences that simulate user requests using ChatGPT, along with 10,000 YouTube comments. For each of the 100 query sentences, the study manually creates three Positive samples, two easy negative samples, and one hard negative sample. Triplets are constructed from all combinations of each query with their corresponding positive and negative samples, resulting in a total of 900 Triplets ($100 \times 3 \times 3$), which are used for Triplet Loss learning.

3.3. Outputs in Recommendations

The proposed method regards recommendation ranking as a classification task. It also indicates the weight of the attention mechanism during classification.

For the classification learning depicted in Figure 3.1, the BERT weights pre-trained using Triplet Loss are used as initial parameters. The model is trained for a music recommendation classification task. As training data, 10,000 YouTube comments are collected from music videos corresponding to the top 100 songs in the 2024 Karaoke DAM ranking[17]. For each song, 100 comments are selected in descending order of the number of preference marks.

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

The model is trained for a classification task with the comments and the song titles as input features and target labels, respectively. The comments are converted into embedded vectors using BERT. Each song title is assigned a unique ID from 1 to 100 to be used as a class label. Through this training process, when a query sentence is provided, the model presents the song with the highest predicted probability as the recommendation result.

In addition to presenting recommended songs, the method visualizes outputs of the attention mechanism. Attention represents weights that indicate how strongly each word in a sentence is referenced with respect to a specific token. The word and the token correspond to a key and a query, respectively. The weights are computed using the SoftMax function. Accordingly, each attention value ranges from 0 to 1. It forms a probability distribution whose sum over the key dimension equals 1 for each query token. The study utilizes the self-attention in the final layer of BERT. Attention weights obtained from multiple heads are averaged to compute the attention score for each word. By visualizing the values as a heatmap, we can intuitively understand which words the model focuses on during recommendation, thereby clarifying the rationale behind the recommendation. It contributes to addressing the black-box nature of conventional recommendation methods.

3.4. Attention Mechanism reflecting individual preferences

Users unsatisfied with the recommendation results can request other results. At that time, they specify any number of words they wish to emphasize or de-emphasize, viewing the recommendation results and attention. In the former case, the attention weights are increased, while they are decreased in the latter. The feedback causes the model to modify the attention weights. They are modified with Triplet Loss learning, similar to the prior nuance learning. In the Triplet Loss retraining process, triplets are constructed using the user query as the anchor. Sentences containing words whose attention is to be increased are used as positive samples, whereas sentences containing words whose attention is to be decreased are used as negative samples. A total of 10 Triplets are created to be used for retraining. Through the process, the model's attention is adjusted in the direction specified by the user. It enables recommendations to better align with user intent, which provides a more satisfying recommendation experience.

4. Attention Differences Based on Learning Presence

4.1. Experimental Details

The study conducts an experiment to evaluate the effectiveness of nuance learning and classification learning. The experiment examines differences in attention among three models for the same query sentence from a user. It prepares three variants: SimpleBERT, standing for a BERT model without fine-tuning, ClassBERT, standing for a BERT model trained only with classification learning, and Triplet-ClassBERT, standing for the proposed model trained with both Triplet Loss–based nuance learning and classification learning. For SimpleBERT, the experiment employs the cl-tohoku/bert-base-japanese model[18], which is widely used for Japanese natural language processing.

4.2. Results and Discussion

Figure 4.1 shows the attention of the three models for the sentence "A song with calm sections and a powerful chorus". The thicker the color, the stronger the attention. All queries and comments used in the study are originally written in Japanese. All figures and tables that appear below are English translations of the original. The attention of Simple BERT focuses on individual words in the query sentence. It instead concentrates on the [SEP] token. The result

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

indicates that SimpleBERT does not sufficiently capture the semantic content of the query sentence. It fails to identify words that are important for recommendations. In contrast, ClassBERT gives high attention values to all independent words in the query sentence. It suggests that ClassBERT can attend to individual words during recommendation unlike SimpleBERT. However, attention is strongly directed not only toward important recommendation words like "quiet," "rust," and "peak," but also toward meaningless words like 'part' that are ineffective for the classification task. It reveals that the attention remains dispersed, making the model insufficient for effective recommendations. Finally, similar to ClassBERT, Triplet-ClassBERT directs its strong attention only to meaningful words such as "quiet," "refrain," and "climax." Furthermore, it pays little attention to other meaningful words. Triplet-ClassBERT focuses more effectively on words crucial for recommendations. The results demonstrate that Triplet-ClassBERT can select words that are more critical for recommendations.

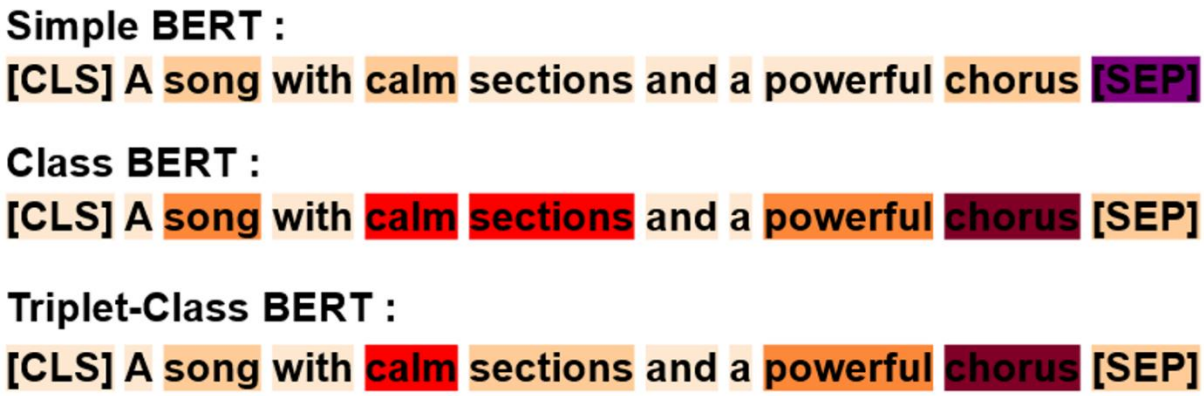

Figure 4.1: Result of attention differences based on learning presence

5. Distinguishing Similar Queries through

5.1. Experimental Details

Another experiment takes place to evaluate the usefulness of nuance learning, which is a component of the proposed method, by examining both attention patterns and recommendation results. The experiment compares ClassBERT and Triplet-ClassBERT when processing two query sentences that vary slightly in content. Their difference in attention patterns and recommendation results demonstrates the effectiveness of nuance learning.

5.2. Results and Discussion

Figure 5.1 and Table 5.1 show the attention patterns and recommendation results of the two models for the two similar query sentences, denoted as queries α and β. In Table 5.1, red indicates songs with positive lyrics or melodies, while blue indicates songs expressing regret. The purple shows songs related to past memories. The scores represent classification probabilities. Based on the query content, red songs should be recommended for query α, while purple songs for query β, followed by blue ones.

In Figure 5.1, for query α, ClassBERT gives strong attention to important content words such as "heartbreak" and "cheerful." However, for the slightly different query β, the strong attention of the model is imposed only on "heartbreak." The model fails to focus on another important word, namely "happy." The results for ClassBERT show that songs with high classification probabilities are almost the same between the two queries. In addition, blue songs dominate the

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

classification probabilities for both queries. Especially for query α, inappropriate songs are ranked highly. These results indicate that ClassBERT focuses strongly only on “heartbreak.” The model fails to grasp the similarity in meaning of ‘cheer’ and “happy memories” in the two queries. It prevents the model from reflecting the similarity in the recommendation results.

Next, let us see Triplet-ClassBERT in Figure 5.1. The model focuses on important words like “heartbreak” and “cheerful,” similar to ClassBERT. However, for query β, its attention is strongly focused on the parts “heartbreak” and “enjoyable,” showing it focuses on important words for both queries. Additionally, Table 5.1 shows that Triplet-ClassBERT recommends a red song for query α, unlike ClassBERT. For query β, the number of blue songs increases, while red ones join. These findings indicate that Triplet-ClassBERT understands not only the element of “heartbreak” but also grasps the nuanced differences between uplifting songs and songs that immerse you in happy memories.

ThereforeThe result shows, through nuance learning via Triplet Loss learning, the model acquires “human-like sensibilities” unattainable through conventional training. The proposed method demonstratesing its ability to modify attention in an ideal manner.

Figure 5.1: Attention results to similar queries for each method

Table 5.1: Recommendation results to similar queries for each method

	ClassBERT				Triplet-ClassBERT			
	Query α		Query β		Query α		Query β	
1	最後の雨(Saigo no Ame)	0.95	最後の雨(Saigo no Ame)	0.70	サウダージ(Saudade)	0.19	サウダージ(Saudade)	0.19
2	花束の代わりにメロディーを(Hanataba no Kawarini Melody o)	0.01	366日(366 nichi)	0.13	チェリー(Cherry)	0.16	ドライフラワー(Dry Flower)	0.19
3	366日(366 nichi)	0.01	Pretender(Pretender)	0.05	最後の雨(Saigo no Ame)	0.13	最後の雨(Saigo no Ame)	0.17
4	Pretender(Pretender)	0.00	カブトムシ(Kabutomushi)	0.01	First Love(First Love)	0.12	First Love(First Love)	0.15
5	ビリミリオン(Birimirion)	0.00	ランデヴー(Randezvous)	0.01	ドライフラワー(Dry Flower)	0.06	カブトムシ(Kabutomushi)	0.06
6	I LOVE YOU(I Love You)	0.00	サウダージ(Saudade)	0.01	Pretender(Pretender)	0.03	チェリー(Cherry)	0.02
7	カブトムシ(Kabutomushi)	0.00	花束の代わりにメロディーを(Hanataba no Kawarini Melody o)	0.01	ビリミリオン(Birimirion)	0.02	シンデレラボーイ(Cinderella Boy)	0.02
8	シングルベッド(Single Bed)	0.00	恋人ごっこ()Koibito Gokko	0.01	シャルル(Charles)	0.02	Pretender(Pretender)	0.02
9	チェリー(Cherry)	0.00	糸(Ito)	0.00	高嶺の花子さん(Takane no Hanako san)	0.02	366日(366 nichi)	0.02
10	世界が終わるまでは…(Sekai ga Owaru made wa…)	0.00	ベテルギウス(Betelgeuse)	0.00	小さな恋の歌(Chiisana Koi no Uta)	0.02	アイノカタチ(Ai no Katachi)	0.02

6. Human Evaluation Experiment

6.1. Experimental Details

The study verifies the overall usefulness of the proposed method and the validity of modifying attention through Triplet Loss learning with an experiment by humans. The experiment provides users with a recommendation experience. The experiment aims to compare user evaluations for attention change before and after the modification for ClassBERT and Triplet-ClassBERT.

The experimental flow is as follows.

1. The user presents a query

2. The top 5 recommended songs from each of ClassBERT and Triplet-ClassBERT for that query are displayed along with Triplet-ClassBERT's attention.

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

3. The user rates their satisfaction with recommended songs by each model on a scale of 1 to 5

4. Based on Triplet-ClassBERT attention, the user specifies words on which they want more/less attention.

5. Triplet-ClassBERT is trained again using the method based on the user's feedback.

6. Each model recommends the top 5 songs again.

7. The user rates the satisfaction with the newly recommended songs on a scale of 1 to 5.

Twenty subjects have participated in the experiment, with each subject generating five types of query sentences, resulting in a total of 100 query sentences used for the experiment.

Note that there are three models in the above flow. The first is ClassBERT without triplets, which is referred to as Model 1. The second and the third are both Triplet-ClassBERT. The second is the one before training with user feedback. It is referred to as Model 2. The last is Triplet-ClassBERT after training with feedback, which is referred to as Model 3.

6.2. Results and Discussion

Table 6.1 summarizes the satisfaction levels for each model. The right-side column 'overall' represents the mean satisfaction across all rankings from the 1st to the 5th place, while each of the remaining columns shows the average rating for each specific ranking position per model. Model 3 achieves higher satisfaction across all rankings compared to Model 2 and Model 2. Furthermore, it receives an exceptionally high rating of 4.15 for the top recommendation probability. These findings indicate that the model refinement via Triplet Loss learning significantly contributes to increased user satisfaction.

Model 2 outperforms Model 1 in the ‘overall’ column. Though it scores lower in 1st and 5th places, it achieves significantly higher ratings than Model 1 for 2nd, 3rd, and 4th place. It indicates that Model 2 possesses the stability of mid-tier items. It can make broadly valid recommendations. The difference of Model 2 from Model 1 occurs due to the nuance learning enabled by Triplet Loss learning. Through nuance learning, the query text and its semantic representation are subdivided, constructing an embedded space that reflects multiple elements contained within the query such as emotion, situation, and atmosphere. Consequently, rather than relying solely on comments containing expressions identical to the query, the model can go beyond exact lexical matches. This enables the capture of comments with similar nuances, such as sentiment or atmosphere, even when they are described using different vocabulary or phrasing within the same semantic space. This explains why Model 2 does not strongly dominate the top-ranked position, while consistently providing plausible recommendations from the second to fourth ranks.

Next, Figure 6.1, Table 6.2, and Figure 6.2 show actual user query and feedback, positive Triplet examples used for retraining, and attention before and after retraining, respectively. Table 6.3 presents the recommendation results and satisfaction rate for each model. They are obtained in the case where the user provides the query “Songs to listen to on winter nights”.

Let us examine the concrete interaction between models and users to discuss the experimental results. The user is presented with the results from Model 1 and Model 2, along with the attention of Model 2. The user specifies words to be strengthened and ones to be weakened, which triggers the retraining of Model 3. As described in Section 3.4, the user query “Songs to listen to on winter nights” is used as the anchor to create the triplet. From 10,000 YouTube comments, the experiment has retrieved ten examples containing the words specified by the user. They are positive and negative examples. Table 6.2 shows a portion of the positive examples. For the examples in Table 6.3, the user specifies “winter” and “night” as the words

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

they want to strengthen based on the attention before training, while the user specifies “song” as a word to be weakened.

Table 6.1: Average satisfaction rating for each model

	Model 1	Model 2	Model 3
Overall	3.17	3.25	3.61
1st	3.58	3.52	4.15
2nd	3.16	3.44	3.56
3rd	3.12	3.46	3.51
4th	2.98	3.06	3.54
5th	3.02	2.78	3.26

Words to emphasize	winter,nights
Words to ignore	Songs

Query: Songs to Listen to on winter nights

Figure 6.1: Example of user query and feedback

Regarding changes in attention before and after retraining, Figure 6.2 shows that attention for “winter” is so high that little change is observed even before retraining. However, attention for ‘night’, the other word to strengthen, and “song,” the word to weaken, has changed effectively as the user requests. It indicates that retraining using Triplet Loss learning successfully changed attention as intended.

Next, let us examine the achievement of the retraining. Table 6.4 shows the success rate of attention changes across the entire experiment. For “All,” this confirms attention for the target and non-target words across all 100 queries given in the experiment. The attention is considered successful if it changed in the desired direction, The table shows the percentage of successes. The numbers in parentheses indicate the total number of words specified in the feedback and the number of words where the change is successful. For example, for reinforcement, 88 words to strengthen are provided overall, and 56 of them, or 63.6%, are successful. The (0.9, 0.1) column represents the success rate narrowed down to only words with pre-relearning attention values of 0.9 or lower to be strengthened, while only words with pre-relearning attention values of 0.1 or higher to be weakened. In this case, for the strengthening, words with an original attention value greater than 0.9 are ignored, while words with an original attention value less than 0.1 are ignored for the weakening. It is considered difficult to further increase large values or decrease small values—in other words, to further radicalize attention. As we move to the right side in Table 6.4, the threshold gets lower. The right side, (0.5, 0.5), means the strengthening is examined only for words with pre-relearning attention ≤ 0.5, while the Weakening Success Rate is examined only for words with pre-relearning attention ≥ 0.5.

Table 6.4 shows that both the strengthening and the weakening have achieved modification success rates of 60–70% overall. Furthermore, the modification success rate increases as the threshold decreases. At the most lenient threshold in Table 6.4, (0.5, 0.5), both the strengthening and the weakening have achieved a very high success rate of approximately 80%. It reveals that the retraining using Triplet Loss learning is quite effective for modifying the proposed attention mechanism, specifically for shifting focus to words previously overlooked or slightly ignoring

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

Table 6.2: Examples of positive example used for Triplet Loss Retraining

positive
The No. 1 song I want to listen to while walking alone on a winter night
Walking down a cold night street while listening to this song through my earbuds. I love this bittersweet feeling of winter.
I really love the feel of winter night air—or maybe it's the smell. That indescribable bittersweet feeling.
Walking home alone on a winter night, I want to listen to backnumber while remembering how I got to talk to the person I like today.
Today, the person I like was humming this song, so I came back to listen again. Winter's almost here, huh.

Figure 6.2: Changes in attention before and after retraining

words that received excessive attention. The overall human experiments confirmed that the sequence of processes demonstrated in this study—pre-training nuance learning via Triplet Loss learning, model white-boxing through attention visualization, obtaining feedback based on this, and model modification via Triplet Loss learning—effectively reduces the gap between user intent and model judgments. This enables the realization of an adjustable music recommendation system that provides convincing explanations for recommendation reasons.

Table 6.3: Examples of recommendation results and satisfaction ratings for each model

Model 1			
Rank	Title	P	Satisfaction(1-5)
1	クリスマスソング(Christmas Song)	0.6661	4
2	瞬き(Mabataki)	0.296	3
3	まちがいさがし(Machigaisagashi)	0.0085	3
4	花束(Hanataba)	0.0053	3
5	アイネクライネ(Aine Kuraine)	0.0043	3

Model 2			
Rank	Title	P	Satisfaction(1-5)
1	クリスマスソング(Christmas Song)	0.9982	4
2	瞬き(Mabataki)	0.0008	3
3	花火(Hanabi)	0.0002	1
4	シャルル(Charles)	0.0001	2
5	さよならエレジー(Sayonara Elegy)	0.0001	3

Model 3 (After Retraining)			
Rank	Title	P	Satisfaction(1-5)
1	クリスマスソング(Christmas Song)	0.8722	5
2	ヒロイン(Heroine)	0.0894	5
3	瞬き(Mabataki)	0.0087	4
4	ヒカリへ(Hikari e)	0.0028	4
5	ハッピーエンド(Happy End)	0.0026	4

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

Table 6.4: Success rate of attention changes by threshold

	All	(0.5,0.5)	(0.6,0.4)	(0.7,0.3)	(0.8,0.2)	(0.9,0.1)
Reinforcement Success Rate(%)	63.6(56/88)	78.8(26/33)	75.6(31/41)	73.1(38/52)	73.2(41/56)	70.5(43/61)
Weakening Success Rate(%)	71.4(55/77)	84.4(38/45)	86.3(44/51)	83.1(49/59)	78.8(52/66)	77.0(57/74)

7. LIMITATION

While the study demonstrates the effectiveness of the proposed method, it has also identified several challenges and limitations that must be addressed for practical implementation.

7.1. Labor

The study demonstrates that model retraining, as presented in Chapter 3, offers significant advantages as a method for modifying models based on user feedback. It addresses issues such as accuracy degradation due to prompt redundancy[19] and the difficulty to address pinpoint changes[20] that often occur with prompt engineering-based approaches. However, it also carries the inevitable disadvantage of computational cost caused by retraining the model. Currently, the proposed method requires approximately 10 to 15 minutes per re-training session. It makes achieving interactive systems difficult, necessitating improvements to the learning method.

LoRA[21] helps by memorizing parts of the model, enabling rapid updates to specific components. The “Classification Learning” step in Figure 3.1 is performed again after the Triplet Loss retraining. LoRA that memorizes the section enables us to modify only the Triplet Loss retraining part. We can maintain the model modification capability while enabling faster recommendations. It brings us closer to achieving an interactive system.

7.2. The Radicalization of Attention

Section 6.2 reveals that retraining using Triplet Loss is quite effective for modifying the proposed pinpoint attention approach—specifically, to focus on words that are previously overlooked or slightly ignore words that have been overemphasized. However, it is also true that success rates decline significantly as the threshold becomes stricter. Significant challenges remain in achieving attention radicalization, such as further emphasizing words that have been already emphasized or further ignoring words that have been already ignored—in other words, further skewing an already biased attention distribution.

The dataset used for retraining via Triplet Loss learning in the method involves keyword searches for comments containing user-specified words. The method creates positive and negative examples from them, to be combined with the user query as the anchor. However, the approach likely hinders effective changes to attention through Triplet Loss learning. It may caused by sentences containing the target words but having vastly different meanings. The target words becoming diluted within longer sentences. To solve the issue, we need to either modify the Triplet data creation method or consider alternative learning approaches beyond Triplet Loss learning.

For instance, we could take into account embedding all sentences, using comments with a certain minimum similarity to the query or specified modification word. Additionally, when creating Triplets, swapping positive and negative examples at a fixed rate might sometimes lead to learning a better embedding space [22]. Introduction of such techniques is a future work.

7.3. Dependency on the field

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

The explainable and model-modifiable system proposed in the study is highly versatile and useful beyond music recommendation. Suppose domain-specific question-answering generation systems[23][24]. The fine-tuning using past questions and their answers, followed by retraining with Triplet Loss, enables deeper and more accurate understanding of user intent. It facilitates the output of more helpful answers, promising faster problem resolution.

However, to realize such systems, it is necessary to develop domain-specific learning methods. Each domain possesses unique phrasing, terminology, and nuances. Some form of nuance learning is required to enable the model to grasp them [25]. Furthermore, regarding model modification methods, it remains unclear whether the approach used in the research will function effectively in other domains. Methods suited to each specific domain must be explored.

8. CONCLUSION

The study focuses on the recognition gap problem in music recommendation. It proposes a recommendation method that not only provides explainable reasons for recommendations but also can be modified based on user feedback. The proposed method consists of the following steps: (1) pre-training for nuance learning using Triplet Loss, (2) presentation of recommendation reasons through attention visualization, and (3) model modification via Triplet Loss re-training based on feedback. Experimental results have confirmed that attention distributions vary depending on the learning method, with the proposed approach showing a tendency for attention to concentrate on words crucial for recommendations. Furthermore, the modification of attention according to user specifications has succeeded approximately 60-70% of the time overall, improving to about 80% when conditions were relaxed. Human evaluations have demonstrated the effectiveness of the proposed method, showing high satisfaction with top recommendations (e.g., average 4.15 for the top recommendation) after retraining.
However, challenges have also been identified, including the time required for retraining, the difficulty of attention radicalization, and domain dependency. Future works aim to accelerate retraining and improve triplet data generation, paving the way for more practical and versatile interactive recommendation systems.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Professor Hiromitsu Shimakawa of Ritsumeikan University for his invaluable guidance in conducting this research and writing this paper. I would also like to thank the members of the Data Engineering Laboratory, College of Information Science and Technology, Ritsumeikan University, who provided me with much advice on my research. I would also like to thank all those who cooperated in the experiment. w

CONFLICT OF INTEREST

The authors declare no conflicts of interest in relation to the publication of this article.

References

[1] Spotify. https://www.spotify.com/. Music streaming service, Accessed: 2025 01-10. Another reference
[2] Apple music. https://www.apple.com/apple-music/. Music streaming service, Accessed: 2025-01-10.
[3] Marta Moscati, Christian Wallmann, Markus Reiter-Haas, Dominik Kowald, Elisabeth Lex, and Markus Schedl. Integrating the act-r framework with collaborative filtering for

Technium Social Sciences Journal
Vol. 81, 67-81, March, 2026
ISSN: 2668-7798
www.techniumscience.com

explainable sequential music recommendation. In Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, p. 840– 847, New York, NY, USA, 2023. Association for Computing Machinery.
[4] MKeesara Sravanthi, G. Yaswanth, M. Manihaar, P. Venkat, and P. Karthik. Mood based music recommendation system. In Frank M. Lin, Ashokkumar Patel, Nishtha Kesswani, and Bosubabu Sambana, editors, Accelerating Discoveries in Data Science and Artificial Intelligence II, pp. 3–10, Cham, 2024. Springer Nature Switzerland.
[5] Nakka Venkata Durga Malleswari, Kuchipudi Gayatri, Katepalli Yaswanth Sai Kumar, Nekkanti Likhita, Padmanaban K, and Debnath Bhattacharyya. Music recommendation system using hybrid approach. In 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), pp. 1560–1564, 2023.
[6] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access, Vol. 6, pp. 52138–52160, 2018.
[7] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019.
[8] Zhikang Dong, Bin Chen, Xiulong Liu, Pawel Polak, and Peng Zhang. Musechat: A conversational music recommendation system for videos, 2024.
[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019.
[10] Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, November 2022. Accessed: 2025-01-10.
[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N.Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.
[12] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models. ACM Computing Surveys (CSUR), Vol. 51, pp. 1– 42, 2018.
[13] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823, 2015.
[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification, 2017.
[15] Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. Hard negative examples are hard, but useful, 2021.
[16] Youtube. https://www.youtube.com. Video sharing platform, Accessed: 2025-01-10.
[17] Karaoke Ranking Top 100 (Latest / Weekly / Monthly). https://mora.jp/topics/news/karaoke2410/. Accessed: 2025-01-10.
[18] cl-tohoku/bert-base-japanese. https://huggingface.co/cl-tohoku/bert-base-japanese. Pretrained Japanese BERT model, Accessed: 2025-01-10.
[19] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, Vol. 55, pp. 1– 35, 2021.
[20] Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. Towards llms robustness to changes in prompt format styles. In North American Chapter of the Association for Computational Linguistics, 2025.
[21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv

preprint arXiv:2106.09685, 2021.
[22] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjoern Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance indeep metric learning. ArXiv, Vol. abs/2002.08473, 2020.
[23] Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rockt¨aschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. ArXiv, Vol. abs/2005.11401, 2020.
[24] Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. Openmedlm: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. Scientific Reports, Vol. 14, 2024.
[25] Suchin Gururangan, Ana Marasovi´c, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don’t stop pretraining: Adapt language models to domains and tasks. ArXiv, Vol. abs/2004.10964, 2020.