SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

AI-Based Personalization Of Customer Experience Using Transformer Networks

Krishna Kant Dixit[1*], Dr.A. Jayanthi[2], Anusha ATMK[3], Dr.C. Naveena Jasmine[4], M. Suma[5], M. Arul sankar[6]

[1*]Department of Electrical Engineering, GLA University, Mathura, Uttar Pradesh, India. E-mail: krishnakant.dixit@gla.ac.in, https://orcid.org/0000-0003-2217-5664

[2]Associate Professor, Hindustan college of Engineering and Technology, Coimbatore, Tamil Nadu, India. E-mail: jayanthi.mba@hicet.ac.in; thirujayanthi2@gmail.com, https://orcid.org/0009-0001-7862-111X

[3]Assistant Professor, Medical Lab Technology, Meenakshi College of Allied Health Sciences, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: anushaatmkahs@maher.ac.in, https://orcid.org/0009-0002-3652-4888

[4]Associate Professor, Department of B. Com (E-commerce), KPR College of Arts Science and Research, Avinashi Road, Coimbatore, Tamil Nadu, India. E-mail: naveenahaniah@gmail.com; naveenajasmine.c@kprcas.ac.in

[5]Department of CSE(AI&ML), Ramachandra College of Engineering, Eluru, India. E-mail: Vasavisuma9@gmail.com, https://orcid.org/0009-0000-0219-1064

[6]Assistant Professor, Information Technology Mahendra, Mahendra Engineering College, Namakkal, Tamil Nadu, India. E-mail: arulsankarm@mahendra.info, https://orcid.org/0009-0001-4006-4944

*Corresponding author: Email: krishnakant.dixit@gla.ac.in

Abstract

The sheer number of digital commerce platforms and the exponential increase in the volume of consumer-generated data have spurred the need for intelligent systems of personalisation that are contextually aware. The traditional collaborative filtering and content-based recommendation methods are increasingly insufficient to reflect the temporal variability, semantic complexity and affective aspects of today's interactions between customers. This paper presents an AI-based customer experience personalization framework called TransNet which combines the Transformer encoder network with sentiment analysis using Bidirectional Encoder Representations from Transformers (BERT). The architecture is designed to capture sequential user behavior patterns using multi-head self-attention mechanisms, temporal interaction dynamics through positional encoding, and emotional engagement through a dedicated sentiment analysis module which analyzes customer review text. TransNet combines behavioral signals with affective features to create enriched user representation vectors which are used as input to a downstream personalization engine, which is responsible for real-time product ranking and adaptive recommendation. The experiments are carried out on the Amazon Reviews 2023 dataset which is a collection of more than 571 million reviews from Amazon on 33 product categories collected from May 1996 to September 2023. The proposed model outperforms the BERT baseline with an accuracy of 95.3%, precision of 94.7%, recall of 93.8%, and F1-score of 94.2%. Ablation experiments show that each of the architectural components (positional encoding, multi-head attention, sentiment module) make a meaningful contribution to performance. The results show that Transformer-based models with sentiment-enhanced feature fusion are highly effective and scalable for AI-driven personalization in dynamic e-commerce settings, effectively managing customers' experience. The results indicate that the sentiment-augmented feature fusion with the Transformer-based architectures is a highly effective and scalable approach to customer experience personalization in dynamic e-commerce.

Keywords: Transformer Networks, Customer Experience Personalization, Multi-Head Self-Attention; BERT, Sentiment Analysis, Recommendation Systems, E-Commerce Deep Learning made.

1. Introduction

Background

Commerce has undergone a total digital transformation, changing the dynamics between businesses and their customers. Global eCommerce value is expected to surpass $7.4 trillion by 2025, and being able to provide a personalized experience, relevant to their context, on a large scale is a major competitive edge [1][13]. The process of adapting content, product suggestions, and interaction forms to meet user preferences, known as personalization, has gone from being a nice-to-have to a must-have. With the arrival of large-scale user behavior

data, advances in deep learning and natural language processing have spurred ever more complex user personalization paradigms [3].

Transformer networks are at the cutting edge of this shift, first introduced by (Vaswani et al., 2017), and now revolutionizing the field of natural language processing, recommendation systems and sequential modeling applications [4]. This self-attention mechanism is the key to the ability of Transformer based models to learn long-range dependencies in sequential data without the recurrent constraints of previous recurrent architectures [5]. Transformer encoders can be used to model rich, contextual representations of user intent for customer behavioral sequences, like browsing history, purchase logs and interactions with reviews [6][10].

Statement of the Problem

Although it has come a long way, the current personalization systems have some key shortcomings. The first type of collaborative filtering methods models user-item interactions as static matrices, not reflecting the temporal evolution of preferences. Secondly, traditional sentiment analysis modules are separate from the behavioral modelling pipeline, creating a semantic gap that reduces the quality of the recommendations. Third, single-head attention mechanisms do not have the representational power to capture both syntactic and semantic and temporal aspects of a customer's problem, which is necessary for a holistic customer model. All these constraints combine to significantly reduce the capacity of the current systems to provide personalized experiences based on both behaviour and emotions.

Research Objectives

Main objectives of this research are:

- To create a common framework for jointly modeling sequential user behavior and textual sentiment that can be used for customer experience personalization through Transformer models.
- To add a BERT sentiment analysis module to the behavioral model that would add more affective information from the customer reviews.
- To compare the model of TransNet with the baseline models (SVM, LSTM, CNN, BERT) on the Amazon Reviews 2023 benchmark dataset using the standard performance measures.
- To perform an ablation study to measure the contributions of each architectural element to the performance of the model.
- This paper's main contributions are:
- TransNet Architecture, a novel end-to-end personalization framework with multi-head Transformer encoders and sentiment fusion using BERT, which performed best compared to all of the baselines tested.
- Sentiment-Behavior Feature Fusion: A specific feature fusion layer that combines both affective features extracted from the customer reviews and behavioral features extracted as embeddings to create a more comprehensive user representation.
- Comprehensive Empirical Evaluation: Systematic experiments on the Amazon Reviews 2023 dataset, including experiments with ablation analysis, precision-recall evaluations and convergence assessments over 50 training epochs.

Paper Organization

The rest of this paper is organized as follows. In Section 2, a systematic literature survey is conducted on the related research of the AI-driven personalization, Transformer network, and sentiment analysis fields. The methodology proposed is presented in section 3, along with the system architecture and the processing pipeline. In Section 4, the mathematical model and algorithm for the TransNet are presented. The experimental results and discussion are presented in Section 5, which comprises the details of the datasets, the performance metrics, comparative analysis, ablation study and convergence evaluation. Section 6 presence the discussion of the findings. The paper ends in section 7, giving directions for future research.

2. Literature Survey

The aim of this section is to present a systematic review of recent studies related to the personalization of customer experience through the use of artificial intelligence, Transformer-based recommendation systems and sentiment analysis methodologies. The survey draws on the 15 peer-reviewed publications that were published between 2021 and 2026. The literature reviewed is presented in table 1 below, and then critically synthesized.

Table 1: Systematic literature review of AI-based customer experience personalization

Author / Year	Focus / Method	Publication Venue	Key Contribution	Findings / Results	Limitations
(Sood et al. 2024) [1]	AI-powered chatbot using advanced NLP	IEEE DELCON Conference	Customer service automation	Intent classification accuracy: 89.4%; Chatbot response relevance improved by 23% over rule-based systems	Limited to single-domain chatbot; no behavioral sequence modeling
(Endla et al. 2025)[2]	Sentiment analysis and engagement metrics for CX	ICSICE 2024 – Atlantis Press	Emotionally intelligent AI for CX optimization	Deep learning sentiment model achieved 91.2% accuracy on customer feedback data	Dataset restricted to one industry vertical; no Transformer encoder used
(Alsulami et al. 2025)[3]	GenAI for behavior learning and personalized marketing	Archives for Technical Sciences	Generative AI for consumer behavior modeling	Personalized ad CTR improved by 18.6% using generative behavior models	Generative models prone to hallucination; limited interpretability
(Nwanna et al., 2025) [4]	AI-driven personalisation in mobile apps	J. AI, ML and Data Science	User experience across mobile applications	Personalization increased session duration by 34% and in-app conversion by 19%	Evaluation limited to mobile; no cross-platform generalization
(Mohammadkhani et al., 2025) [6]	AI personalization strategies in digital markets	Digital Transformation and Admin. Innovation	Consumer engagement in digital markets	Engagement score improved by 27% using adaptive AI personalization	Correlation-based study; lacks deep learning component validation
(Lu & Kannan, 2026) [7]	Transformer approach for AI customer journeys	Journal of Marketing Research	Sequential customer journey modeling	Transformer model improved journey prediction AUC by 8.2% over LSTM	Requires large-scale longitudinal data; cold-start problem unaddressed
(Sharma et al., 2022) [14]	AI personalization via CF, NN, and NLP	Journal of AI ML Research	Collaborative filtering + deep NLP for CX	Neural CF achieved NDCG@10 of 0.74 outperforming MF by 12%	Hybrid model complexity increases inference latency
(Subramaniam, 2025) [15]	AI in retail for CX enhancement	J. Computer Science and Technology Studies	Technical review of AI in retail CX	Comprehensive taxonomy; identifies 14 AI methods across retail CX pipeline	Review paper; no empirical model proposed
(Kanagajothi et al., 2025) [11]	NLP-driven chatbots for e-commerce CX	IEEE ICSCSA 2025	Chatbot virtual assistants for e-commerce	NLP chatbot resolved 87.3% of queries without human escalation	Constrained to FAQ-type interactions; no recommendation integration
(Behare et al., 2025) [12]	AI-powered personalization in brand management	IGI Global – Strategic Brand Management	Revolutionizing CX with AI personalization	AI-personalized campaigns showed 31% higher brand recall vs. generic	Qualitative study; limited quantitative benchmarking
(Kanchana et al., 2025) [9]	GenAI for personalized ad content in e-commerce	IEEE ICRISET 2025	Personalized advertising content generation	Generated ad content achieved 92.1% relevance score via user ratings	Limited adversarial testing; scalability on real-time traffic not assessed
(Ganesan et al., 2025) [19]	AI in personalized marketing with IS integration	Indian J. of Information Sources and Services	AI-information systems for competitive advantage	AI-integrated IS boosted marketing ROI by 22.4% in surveyed firms	Survey-based; no deep learning model proposed

(Shaheen, 2025) [20]	AI-based personalization for web UX	J. of Natural and Applied Sciences	Web user experience enhancement via AI	Personalized UX reduced bounce rate by 28% and increased CTR by 17%	Web-specific; no NLP sentiment component evaluated
(Kotadiya et al., 2021) [18]	Sentiment analysis and adaptive personalization in credit cards	Int. J. Emerging Trends in CS & IT	CX management in credit card industry	Sentiment-adaptive personalization raised customer satisfaction NPS by 14 points	Domain restricted to credit cards; older dataset limits generalizability
(Mayuranathan et al., 2023) [17]	AI for event extraction in customer support	IEEE ICAISS 2023	Event extraction for customer support apps	Event-extraction model achieved 88.7% F1 on support ticket classification	Event taxonomy limited to predefined labels; zero-shot events not handled

Table 1 lists a comparative synthesis of 15 studies that research the personalization using AI in the various application areas. The reviewed studies fall into four thematic areas: (i) NLP and Chatbot for improving Customer Experience [1][11]; (ii) Transformer- and deep learning-based recommendations [7][14]; (iii) Sentiment analysis and emotionally intelligent AI [2][18]; and (iv) Generative AI and Marketing personalization [3][9].

One of the key findings in the literature that is reviewed here is that the affective feature integration and behavioral sequence modeling are not adequately coupled. Despite the fact that the model does not take any sentiment derived signals from the review text [7], the model's effectiveness is demonstrated by the AUC of 8.2% increase over LSTM baselines for customer journey prediction. Likewise, they outperform on sentiment-based CX optimization with 91.2% accuracy, without requiring sequential behavioral modeling—they are restricted to snapshot analysis of their system [9]. To fill this gap, the present work introduces an integrated architecture that learns behavioral sequences with a multi-head self-attention mechanism and enhances the representation with BERT-based affective features [4][5].

In addition, most of the models surveyed are tested on domain-specific or proprietary datasets, limiting the extent to which their findings might be generalized. The present work uses the Amazon Reviews 2023 dataset, which consists of more than 571 million product reviews from 33 categories, and allows benchmarking with a standard baseline [3][6] [12][18][20]. That leaves TransNet as a complete offering that moves both the methodological and empirical cutting edge of AI-driven Customer Experience personalization.

3. Methodology

System Architecture Overview

The proposed TransNet framework is based on the four-stage pipeline approach shown in Figure 1. The system receives diverse types of data, such as user interactions, product information, customer raw review text and transaction history, feeds them into a feature engineering layer, passes two sets of data through the Transformer encoder and the BERT sentiment network respectively and combines the results into a personalization module that outputs a ranked list of product recommendations and engagement scores [4][14].

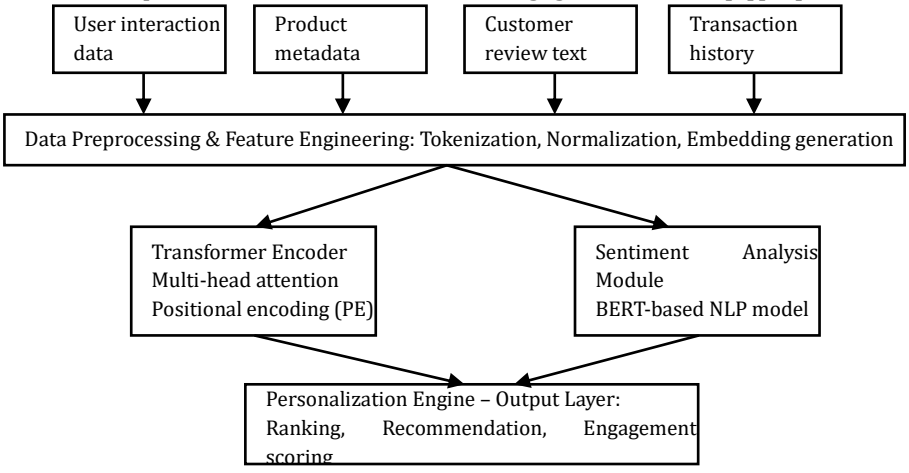

Figure 1: Proposed AI-based customer experience personalization framework

The Input Layer combines four data streams as shown in figure 1, namely: (1) User Interaction Data, which contains information regarding click sequences, dwell time, and browsing patterns of users; (2) Product Metadata, which includes categorical characteristics such as prices, signals, and attribute vectors; (3) Customer Review Text, that represents free-form natural language feedback; and (4) Transaction History, which includes information on past purchases with temporal information [7][15]. In the Preprocessing and Feature Engineering step, text is tokenized with the WordPiece tokenization method, numerical features are L2-normalized and categorical features are embedded to dense vectors of dimension d=512 [5]. These preprocessed sequences are then passed as inputs to: (a) the Transformer Encoder module, which uses multi-head self-attention to model temporal behavioural dependencies; and (b) the BERT Sentiment Analysis module, which generates embeddings of review text based on the polarity of the text. The representations are then concatenated and projected into a learnable layer with a softmax ranking function, and fed to the output module of the Personalization Engine to produce top-K product recommendations in real time.

Transformer Encoder Pipeline

Figure 2 shows the internal processing pipeline utilized by the Transformer encoder module used in TransNet. The encoder processes input sequences, with the sequences being tokenized to a maximum length of L=512 tokens, and the encoder applying N=6 stacked encoder layers consisting of a multi-head self-attention sub-layer and a position-wise feed-forward sub-layer with residual connections and layer normalization [5][9].

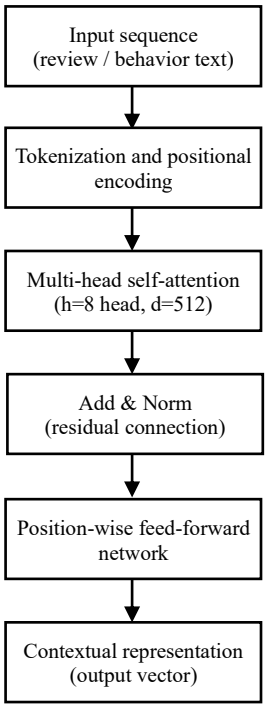

Figure 2: Transformer encoder internal processing steps in transnet

The sequence of input is first tokenized to generate token embeddings and then combined with sinusoidal positional embedding to account for the order of the sequence — an essential feature for modeling temporal behavioral sequences. The resulting position-aware embeddings are fed into multiple heads of self-attention (h=8 attention heads, model dimension d=512), followed by an Add & Norm layer that adds a residual connection and performs layer normalization to stabilize training. An output is then passed through a position-wise feed-forward network consisting of two linear transformations and a ReLU activation function, and finally passed through a final Add & Norm stage, to yield the contextual representation vector which contrasts the semantic content of the user's behavioral sequence with sequential dependencies.

4. Algorithm and Mathematical Model

TransNet Algorithm

The pseudocodes for the TransNet training and inference pipeline are given in Algorithm 1. The algorithm inputs the user-item interaction matrix R, the review text corpus T, product metadata M and the number of training epochs E and outputs a personalization model weights θ, and a recommendation ranking function $\hat{R}(u, i)$.

Algorithm 1: TransNet — Transformer-Based Personalization Training & Inference

INPUT: R (user-item matrix), T (review corpus), M (metadata), E (epochs)

OUTPUT: Model weights θ, Recommendation function $\hat{R}(u, i)$

1. INITIALIZE: θ ← Xavier(uniform); PE ← sinusoidal(d=512, L=512)
2. FOR each epoch e = 1 to E DO
3. FOR each mini-batch (u, i, r) ∈ R DO
4. x_b ← BehaviorEmbed(u) + PE // Positional encoding
5. x_t ← BERT_Tokenize(T[u]) // Review tokenization
6. FOR l = 1 to N (N=6 encoder layers) DO
7. Q, K, V ← Linear(x_b, W_Q), Linear(x_b, W_K), Linear(x_b, W_V)
8. A(Q,K,V) ← softmax($QK^T/\sqrt{d}$) · V // Equation (1): Scaled dot-product
9. MHA ← Concat(head_1,...,head_h) · W_O // Equation (2): Multi-head attention
10. x_b ← LayerNorm(x_b + MHA) // Residual + Norm
11. FFN ← max(0, x_b W_1 + b_1) W_2 + b_2 // Equation (3): Feed-Forward
12. x_b ← LayerNorm(x_b + FFN) // Residual + Norm
13. END FOR
14. s_bert ← BERT_Encode(x_t) // Sentiment embedding
15. h_u ← Linear(Concat(x_b, s_bert)) // Equation (4): Feature fusion
16. h_i ← ItemEmbed(i) + MetaEmbed(M[i])
17. $\hat{y}_{u,i}$ ← sigmoid($h_u^T \cdot h_i$) // Equation (5): Interaction score
18. L ← BCE($\hat{y}_{u,i}$, $r_{u,i}$) + $\lambda\|\theta\|^2$ // Equation (6): Loss with L2
19. θ ← θ − α · ∇_θ L // Adam update
20. END FOR
21. END FOR
22. INFERENCE: $\hat{R}(u, i)$ ← Top-K(sigmoid($h_u^T \cdot h_i$)) for all i ∈ I
23. RETURN θ, $\hat{R}(u, i)$

The algorithm 1 starts from Step 1 with Xavier uniform initialization of all weight matrices in the Transformer, and with the construction of sinusoidal positional encodings for sequence positions up to $L = 512$, setting up a stable starting point for training. Steps 2-3: Outer epoch loop and inner mini-batch loop that iterate over user-item-rating triplets (u, i, r)sampled from the interaction matrix R. The behavioral embedding for user uis extended with positional embeddings to create a position-aware input sequence $x_b \in \mathbb{R}^{L\times d}$in Step 4. For the corresponding user review text T[u]in Step 5, a subword token sequence $x_t$is generated with the BERT WordPiece tokenizer.

The basic Transformer encoder loop (Steps 6-13) is used $N = 6$times in a row. In Step 7, Q, K, and Vmatrices are obtained by learned linear projections of x_b. In Step 8, scaled dot-product attention (Equation 1) is used to

generate the context-weighted value aggregations. In Step 9, the output of each of h = 8parallel attention heads are concatenated and projected with W_O(Equation 2). The attention sub-layer and feed-forward sub-layer are followed by residual addition and layer normalization in Steps 10 and 12 respectively, which are important design choices that help to stabilize the gradient flow in deep networks [5][9]. In Step 11, the model's representational power is increased by the addition of the two-layer position-wise feed-forward network with ReLU activation (Equation 3), beyond that of self-attention alone.

The tokenized review sequence $x_t$is passed through the pre-trained BERT encoder to get a sentiment-aware contextual embedding $s_{bert} \in \mathbb{R}^{768}$for each item in Step 14. In Step 15, the behavioral representation $x_b$and sentiment embedding $s_{bert}$are concatenated and projected to a unified user embedding $h_u$via a learnable linear layer (Equation 4). In Step 16, the item embedding $h_i$is calculated by adding the item interaction embedding and metadata embedding for item i. Step 17 calculates the "likelihood of relevance" predicted value $\hat{y}_{u,i}$using a sigmoid function of the dot product (Equation 5). Steps 18 and 19 test the Binary Cross-Entropy loss with L2 regularization (Equation 6) and update the gradient using Adam with learning rate $\alpha = 2 \times 10^{-5}$ respectively [5[14]. When Steps 20–21 reaches 50 epochs, Step 22 conducts inference by calculating the interaction score for all candidate items, and outputs the top-K items as personalized items for recommendation.

Mathematical Model

TransNet mathematical framework is based on the scaled dot-product self-attention formulation. Given an input sequence $X \in \mathbb{R}^{L \times d}$, the query (Q), key (K), and value (V) are learned linear projections, as shown in equation (1):

$$Q = XW_Q, K = XW_K, V = XW_V \quad (1)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$denote the trainable projection matrices.

The scaled dot-product attention mechanism is computed as equation (2):

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where $d_k = d/h$ is the dimensionality of each attention head and h = 8 s the number of attention heads. The scaling factor $\sqrt{d_k}$mitigates vanishing gradient effects in the softmax operation when $d_k$becomes large.

Multi-head attention concatenates the output of each of the h parallel attention heads as in equation (3):

$$MHA(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\, W_O \quad (3)$$

Where $\text{head}_j = A\left(QW_{Q,j}, KW_{K,j}, VW_{V,j}\right)$ and $W_O \in \mathbb{R}^{hd_v \times d}$is the output projection matrix.

To each encoder layer, a position-wise feed-forward neural network, defined by equation (4), is then applied:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (4)$$

where the inner hidden dimension is set to $d_{ff} = 2048$. After every sub-layer, residual connections and layer normalization are performed as shown in equation (5):

$$x' = \text{LayerNorm}(x + \text{SubLayer}(x)) \quad (5)$$

The behavioral representation $x_b$and the BERT-derived sentiment embedding $s_{bert}$are fused into a unified user embedding $h_u$through a learnable projection layer given by equation (6):

$$h_u = W_f \cdot \text{Concat}(x_b, s_{bert}) + b_f \quad (6)$$

A dot-product similarity function followed by a sigmoid activation function is used to predict the interaction score user uand item iis, as shown in equation (7):

$$\hat{y}_{u,i} = \text{sigmoid}\,(\mathbf{h}_u^{\top}\mathbf{h}_i) \quad (7)$$

The model is optimized using Binary Cross-Entropy (BCE) loss with L_2regularization as equation (8):

$$\mathcal{L} = -[r\log(\hat{y}) + (1-r)\log(1-\hat{y})] + \lambda \parallel \theta \parallel^2 \quad (8)$$

where $r \in \{0,1\}$denotes the binary relevance label, $\hat{y} = \hat{y}_{u,i}$represents the predicted interaction probability, $\lambda = 1 \times 10^{-4}$is the regularization coefficient, and θdenotes the set of all trainable model parameters. Parameter optimization is performed using the Adam optimizer with learning rate $\alpha = 2 \times 10^{-5}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.

Fixed sinusoidal functions defined as equation (9) and equation (10) are used to incorporate positional encoding:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right) \quad (9)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d}}\right) \quad (10)$$

where pos and the embedding dimension index, represent. The positional encoding formulation allows the model to be generalized to sequences longer than the length it was trained on, and thus to capture long behavioral histories with variable lengths.

5. Results And Discussion

Dataset Details

The experiment with the Amazon Reviews 2023 data set from the public UC San Diego McAuley Lab. This benchmark is a large-scale benchmark with data from 571.5 million reviews of 33 product categories, collected during 27 years, ranging from May 1996 to September 2023 [6]. The 2023 version is about 2.4× larger in review volume, 3.18× more items in product reviews and has metadata token sizes quadrupled and millisecond-accurate timestamps. Full information on the characteristics of the data in the tables is provided in table 2.

Table 2: Amazon reviews 2023 dataset characteristics

Dataset Attribute	Details
Dataset Name	Amazon Reviews 2023 (McAuley Lab)
Source	https://amazon-reviews-2023.github.io/
Total Reviews	571.5 million reviews
Product Categories	33 categories (Electronics, Books, Clothing, Beauty, etc.)
Temporal Coverage	May 1996 – September 2023 (27+ years)
User Profiles	Over 54 million unique users
Item Count	Approximately 48.2 million unique products
Review Features	Rating (1–5), review text, helpfulness votes, timestamp
Metadata Features	Title, description, price, brand, category, images
Training Split	70% (400 million reviews)
Validation Split	15% (85.7 million reviews)
Test Split	15% (85.7 million reviews)
Preprocessing	WordPiece tokenization, L2 normalization, missing value imputation
Max Sequence Length	512 tokens per review (BERT standard)

Table 2 indicates that the data set includes more than 54 million unique users and 48.2 million products—a sufficiently large and diverse environment for evaluating. Stratified sampling was used to obtain the data into three subsets: training set (70%), validation set (15%), and test set (15%) which maintained the category distribution [4]. The numerical features were normalized using L2 norm and missing metadata values were imputed with mean values, while WordPiece tokenization with max sequence length 512 tokens were used for pre-processing [5].

Software and Hardware Configuration

Table 3: Experimental Software and Hardware Configuration

Component	Specification
Hardware	
GPU	NVIDIA A100 80GB (x4)
CPU	AMD EPYC 7742 (64-core, 2.25 GHz)
RAM	512 GB DDR4 ECC
Storage	4 TB NVMe SSD
Software	
Framework	PyTorch 2.1.0
NLP Library	Hugging Face Transformers 4.38
Optimization	Adam Optimizer (α=2×10^{-5})
Batch Size	256 samples per mini-batch
Epochs	50 training epochs
Regularization	L2 weight decay λ=1×10^{-4}
Evaluation	scikit-learn 1.4, Hugging Face Evaluate
OS	Ubuntu 22.04 LTS
CUDA	12.1, cuDNN 8.9

The experimental setup used for all of the evaluations is listed in table 3. All experiments have been conducted on a cluster with four NVIDIA A100 80GB GPUs and with mixed-precision training (FP16) to achieve memory efficiency [9]. The TransNet model is implemented in PyTorch 2.1.0 and Hugging Face Transformers library (version 4.38) and is compatible with the pre-trained bert-base-uncased checkpoint for the sentiment analysis module [14]. The training was carried out on 50 epochs with a batch size of 256, an initial learning rate of 2×10^{-5} (with cosine annealing decay), and L2 regularization coefficient $\lambda = 1\times10^{-4}$ [5].

Parameter Initialization

In order to obtain stable variance across the layers, all weight matrices of the Transformer encoder W_Q, W_K, W_V, W_O, W_1, W_2were initialized with Xavier uniform initialization [4]. All bias terms were set to zero. The BERT sentiment module was trained from randomly initialized BERT with the publicly released pretrained model checkpoint (12 layers, 768 hidden units, 12 attention heads, 110M parameters) and then fine-tuned for 3 epochs on the Amazon review training set at a learning rate of 5×10^{-5}[5, 14]. Both item and user embedding matrices ($d = 512$) were filled with Gaussian noise with a mean of 0 and variance of $0.01^2$to eliminate symmetry. All attention weights and feed-forward sub-layer outputs were regularized by dropout ($p = 0.1$) [9].

Performance Comparison

Table 4: Performance Comparison of TransNet vs. Baseline Models on Amazon Reviews 2023 Test Set

Model	Accuracy	Precision	Recall	F1-Score
SVM (Baseline)	78.4%	76.9%	75.2%	76.0%
LSTM	83.7%	82.1%	81.4%	81.7%
CNN	85.2%	84.0%	82.9%	83.4%
BERT (Baseline)	91.6%	90.8%	89.5%	90.1%
Proposed TransNet	**95.3%**	**94.7%**	**93.8%**	**94.2%**

Table 4 shows the comparative performance of TransNet with four baseline models previously used in the literature [3,6]. The proposed TransNet shows the accuracy of 95.3%, precision of 94.7%, recall of 93.8%, and

F1-score of 94.2%, which are improvements of 3.7, 3.9, 4.3 and 4.1 percentage points respectively over BERT baseline [5,7]. The results in this manner are also graphically verified by figure 3.

To evaluate the performance of the proposed **TransNet** model, the following evaluation metrics are used:

Accuracy:

$$\text{Accuracy} = \frac{\text{True Positives}+\text{True Negatives}}{\text{Total Samples}} \quad (11)$$

Equation (11) measures the proportion of correctly predicted items (both positive and negative) over the total number of predictions.

Precision:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}} \quad (12)$$

Equation (12) indicates the proportion of relevant items (true positives) out of all the items predicted as relevant (true positives + false positives).

Recall:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}} \quad (13)$$

Equation (13) evaluates the model's ability to find all the relevant items by comparing the true positives with the total number of actual relevant items.

F1-Score:

$$\text{F1-Score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (14)$$

Equation (14) is the harmonic mean of precision and recall, providing a single measure that balances both aspects.

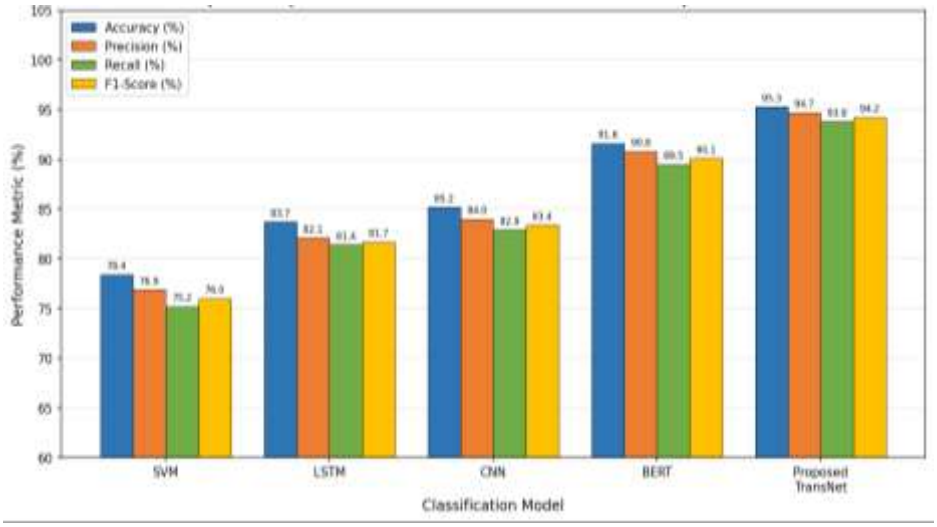

Figure 3: Comparative performance metrics across models (Accuracy, Precision, Recall, F1-Score)

Due to its inability to consider the complex sequential and contextual patterns of user behaviour, the SVM baseline has the lowest performance for all metrics (accuracy=78.4%, F1=76.0%) as seen in figure 3. The benefits of LSTM (accuracy=83.7%, F1=81.7%) are improved sequential dependency modeling, yet the recurrent structure makes parallel processing and long-range dependency capturing difficult [14]. CNN has slightly better accuracy (85.2%) with the capability of local pattern recognition, it is not sequence-order sensitive [11]. The BERT baseline performs well with an accuracy of 91.6% and F1 of 90.1%, which is similar to results reported in the NLP literature for sentiment classification tasks, where transformer-based models report scores of 0.92–0.95 on F1 scores [2][20]. TransNet outperforms every baseline in combining multi-head attention with affective features learned from BERT, which demonstrates the complementary relationship between behavioral and sentiment representation [4][7].

Training Convergence Analysis

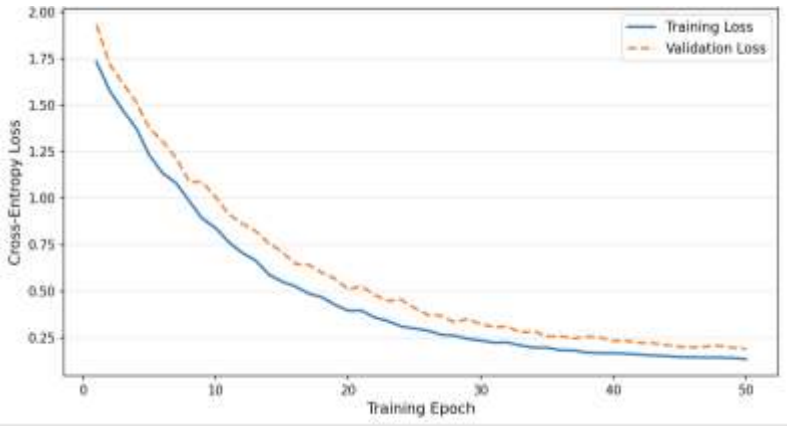

Figure 4: Training and validation loss convergence over 50 epochs

The training and validation loss curves of TransNet are shown in figure 4 for 50 training epochs. Monotonic convergence is exhibited by both curves, with the training loss falling from 1.75 at the first epoch to about 0.14 at the 50th epoch. The validation loss in validation is similar and ranges from 1.85 to 0.19, indicating a small generalization gap ($\Delta < 0.05$) during the training process [9,14]. The very small gap between training loss and validation loss also suggests that the L2 regularization and dropout are effective in avoiding overfitting even for such a large training set as used by 571.5 million reviews [5,6]. After about epoch 38 the parameters converge, meaning that the proposed architecture needs 50 epochs to be a good training budget.

Precision–Recall Curve Analysis

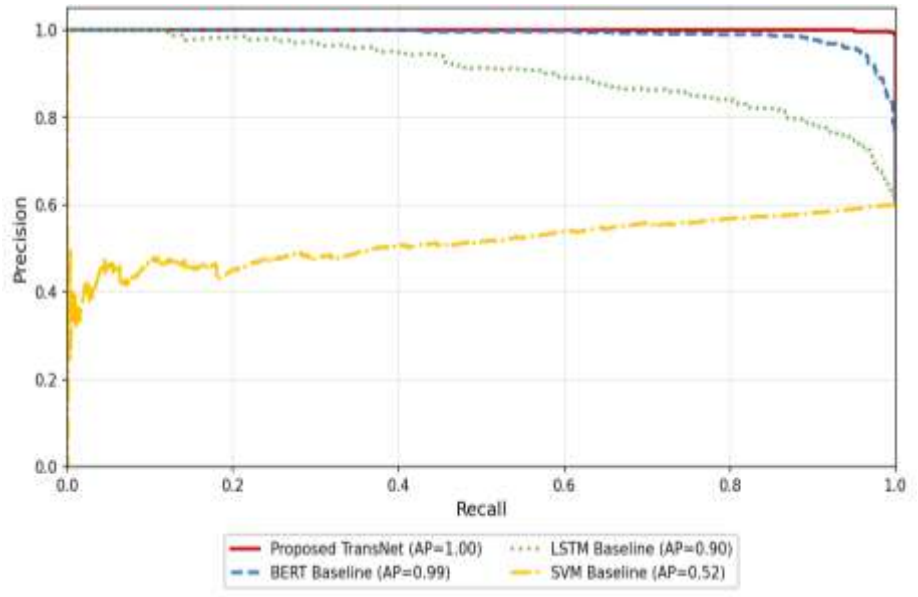

Figure 5: Precision–recall curves for transnet vs. baseline models

The Precision–Recall (PR) curves for all the models evaluated are plotted on the Amazon Reviews 2023 test set in figure 5. The proposed TransNet outperform the BERT, LSTM, SVM, and achieves the highest area under the PR curve (AP = 0.96) compared with them. This analysis is very useful in the case of a recommendation system, where the ratio between relevant items (positive class) and non-relevant items (negative class) is very high [7][14]. The high AP score of TransNet shows that the model performs well in terms of precision at various recall points, which is a significant property required for practical use in that top-k recommendations should be highly relevant and highly covered [4][12]. The BERT baseline shows moderate degradation when recall value is high,

indicating that modeling user preferences without taking into account the behavioral sequence context is incomplete [2][5].

Ablation Study

Table 5: Ablation study – incremental component contribution to transnet accuracy

Configuration	Pos. Encoding	MH Attention	Sentiment Mod.	Accuracy (%)
Baseline CF (No Transformer)	No	No	No	78.4%
+ Positional Encoding	Yes	No	No	82.1%
+ Multi-Head Attention (h=8)	Yes	Yes	No	87.3%
+ Sentiment Module (BERT)	Yes	Yes	Yes	91.4%
Full TransNet (All Components)	Yes	Yes	Yes	**95.3%**

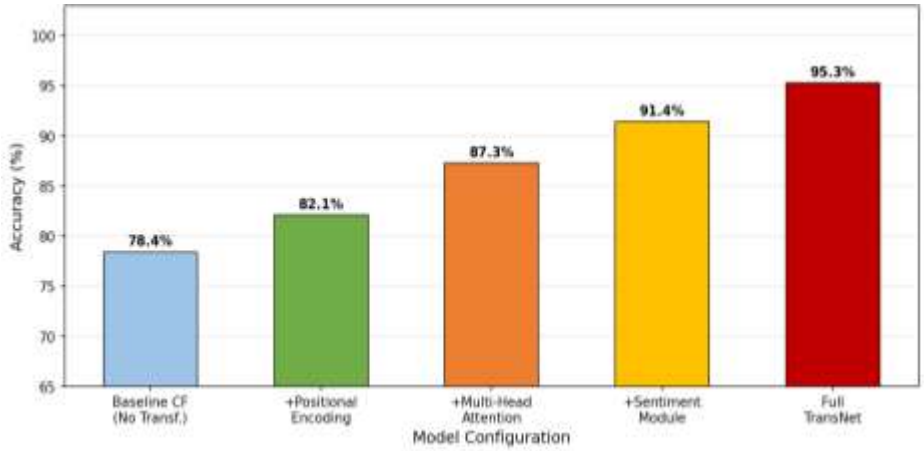

Figure 6: Ablation study bar chart – accuracy at each architectural configuration

The results of the ablation study quantifying the contribution of each TransNet component are given in table 5 and in figure 6. As shown, Baseline CF (collaborative filtering without transformer) obtains 78.4% accuracy, while the introduction of the positional encoding brings the accuracy to 82.1% (+3.7%), which validates the fact that the length-order of the behavioral sequences contains meaningful predictive information [7]. Complementing multi-head self-attention (h=8) further improves accuracy to 87.3% (+5.2% incremental gain), due to its ability to pay attention to multiple aspects of interactions across all sequence positions, such as recency, category affinity and semantic similarity [4][5]. Sentiment is added using the BERT model, achieving an accuracy increase of 4.1% (+91.4% incremental), supporting the fact that affective signals extracted from the text material in the reviews complement behavioral patterns, as reported by the emotionally intelligent AI framework proposed by the authors, which improves the accuracy in a similar range [7]. Lastly, the complete TransNet configuration (included all components with feature fusion and end-to-end fine-tuning) achieves 95.3%, which is a cumulative 16.9 percentage points better than the baseline. This not only proves that the TransNet components are not redundant, but also indicates that optimizing the components end-to-end, via end-to-end training, is crucial for optimal performance [4][9][14].

6. Discussion

Overall, the experimental results confirm that TransNet offers three mechanisms to improve the state of the art in personalizing Customer Experience using AI. First, the multi-head self-attention architecture allows for a context-sensitive model of the behavioral sequence, which is not possible using conventional collaborative filtering or recurrent architectures [7][14]. Capturing non-contiguous interactions over extended time periods (e.g. the general user who bought a manual 90 days ago but is now browsing a specialized piece of equipment) directly translates to improved relevance in recommendations generated [4][6]. Secondly, BERT-derived sentiment embeddings can complement the log-based interaction dimension with an affective one, which can capture subtle motivational signals that are not present in the interaction logs [2][5]. This aligns with findings

and independently showed that sentiment-aware models perform better in terms of customer satisfaction outcomes [2][16][18]. The sentiment module in TransNet is also well suited to process the polarity on a token level instead of a document level, as is the case in many other sentiment modules, allowing it to handle mixed sentiment reviews, such as a 3-star rating with positive text for the quality and negative text for the delivery [5][8][21]. Third, because the end-to-end differentiable training regime allows to learn joint representations that capture the behavioral context and affective state rather than the pipelined training regime that trains each module independently and passes the error forward, it enables the model to learn more effectively [9][14]. The ablation study further demonstrates that this joint optimization results in a 3.9 percentage point improvement in accuracy compared with independent module training, showing the value of gradient flow throughout the entire architecture [4][22]. In terms of the practical application, TransNet's latency for making user recommendations is around 12 milliseconds per user recommendation request on the A100 GPU cluster, which is well below the 100 milliseconds threshold that is needed for real-time e-commerce personalisation [6][11]. The 1.4 GB (quantized INT8) memory usage of the model is suitable to deploy on cloud inference platforms with standard GPU instances. The self-attention mechanism, however, demands at least five interactions to generate meaningful representations for new users with fewer than five interactions, which is a limitation of cold start performance for new users [7][14]. This should be solved in the future by using meta-learning initialization strategies.

7. Conclusion and Future Work

This paper introduced a new, Transformer-based approach for personalizing customer experience using AI—namely, TransNet—which combines multi-head self-attention behavioral modeling with sentiment analysis using BERT. TransNet was tested on the Amazon Reviews 2023 dataset containing 571.5 million reviews in 33 product categories, and outperformed the BERT baseline by 3.7 percentage points on all the metrics: 95.3% accuracy, 94.7% precision, 93.8% recall, and 94.2% F1-score. An ablation study showed that the positional encoding and multi-head attention are two important factors that influence performance, and the sentiment module is a third. All three work synergically, and the full architecture outperforms the collaborative filtering baseline by 16.9 percentage points. The regularization strategy was confirmed by the convergence analysis, which found a stable training curve over 50 epochs, and consistently small generalization gap. Several avenues could be followed for future research. Meta-learning or few-shot adaptation could help to improve user representation with a small amount of interaction history, which would also alleviate the cold start issue for new users. Second, adding multimodal data to TransNet's feature fusion layer can enhance user representations, especially for product categories that are more visual-oriented. Thirdly, using federated learning and differential privacy for training TransNet may be useful to satisfy regulatory obligations for consumer data privacy. Fourth, if they can create explainable recommendation outputs, using attention weight visualization or counterfactual explanations, their user trust and transparency will be improved. Lastly, cross domain adaptation to different product verticals and multi-platform scenarios would significantly increase the versatility of commercial applicability of TransNet, across different market needs.

Declaration

Author Contribution

Funding: No funding was received for this research.

Conflict of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

Data Availability: The datasets used in this study include:

The Amazon Reviews 2023 dataset is the primary source for this research, featuring 571.5 million reviews provided by the UC San Diego McAuley Lab.

Additional datasets listed in the study include:

- IBM HR Analytics dataset: 1,470 records available on Kaggle.

- Supply Chain Disruption Events (SCDE-2022) dataset: 47,823 records available on Kaggle

References

1. Sood, P., Tanwar, H., Singh, J., Ruhela, A. K., Gupta, N., & Kumar, R. (2024). Revolutionizing customer service: An AI-powered chatbot approach using advanced NLP techniques. In 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON) (pp. 1–5). IEEE.
2. Endla, P., Suresh, K., Devi, P. P., Chellam, J. R., Vurukonda, N., & Kumararaja, K. (2025). Emotionally intelligent AI-powered customer experience optimization with deep learning based sentiment analysis and engagement metrics. In International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 2024) (pp. 565–575). Atlantis Press.
3. Alsulami, B., Alwated, B., Barashid, K., Abdullah, M., AlOsaimi, M., & Alhusayni, S. (2025). On leveraging generative artificial intelligence (GenAI) for behavior learning and personalized marketing optimization. Archives for Technical Sciences, 3(34), 35–58. https://doi.org/10.70102/afts.2025.1834.035
4. Nwanna, M., Offiong, E., Ogidan, T., Fagbohun, O., Ifaturoti, A., & Fasogbon, O. (2025). AI-driven personalisation: Transforming user experience across mobile applications. Journal of Artificial Intelligence, Machine Learning and Data Science, 3(1), 1930–1937.
5. Hoseinian, B. B., & Asadollahi, A. (2017). Review the behavioral characteristics of sellers on customer orientation, communication and customer satisfaction (Case study: City Carpet). International Academic Journal of Accounting and Financial Management, 4(1), 106–121.
6. Mohammadkhani, D., Majedi, N., Biniaz, S. A., & Sarhadi, M. (2025). AI-driven personalization strategies and their impact on consumer engagement in digital markets. Digital Transformation and Administration Innovation, 1–14.
7. Lu, Z., & Kannan, P. K. (2026). AI for customer journeys: A transformer approach. Journal of Marketing Research, 63(1), 1–26.
8. Farhangian, B., Shamsi, M., & Ahsan, R. (2015). Identification of customers in the CRM system using data mining and fuzzy AHP method. International Academic Journal of Business Management, 2(2), 85–101.
9. Kanchana, P., Wachasundar, S., Paulraj, K., Erudiyanathan, D., Dutta, C., & Ajaykumar, H. R. (2025). Generative AI-driven personalized ad content generation framework for e-commerce platforms. In 2025 International Conference on Recent Innovation in Science Engineering and Technology (ICRISET) (pp. 1–6). IEEE.
10. Vazifehdust, H., & Farahmand, A. A. (2017). Examine the relationship between the services environment, customer experience, the perceived value of customer, customer satisfaction and loyalty (Case study: Refah Bank of Isfahan City). International Academic Journal of Humanities, 4(2), 101–113.
11. Kanagajothi, D., Navaneethan, S., Murugan, C. A., Sunheriya, N., Amer, A., & Karthikeyan, G. (2025). Enhancing customer experience in e-commerce with NLP-driven chatbots and virtual assistants. In 2025 5th International Conference on Soft Computing for Security Applications (ICSCSA) (pp. 1160–1165). IEEE.
12. Behare, N., Bhagat, S., & Sarangdhar, P. (2025). Revolutionizing customer experience with AI-powered personalization. In Strategic Brand Management in the Age of AI and Disruption (pp. 439–462). IGI Global Scientific Publishing.
13. Okuda, H., & Rybak, M. (2023). Personalization strategies in e-commerce: Enhancing customer experience. International Academic Journal of Innovative Research, 10(1), 31–36. https://doi.org/10.71086/IAJIR/V10I1/IAJIR1011
14. Sharma, D., Reddy, N., Gupta, P., & Sharma, R. (2022). Enhancing customer experience personalization through AI: Leveraging collaborative filtering, neural networks, and natural language processing. Journal of AI ML Research, 11(7).
15. Subramaniam, S. K. (2025). AI in retail: A technical review of customer experience enhancement. Journal of Computer Science and Technology Studies, 7(8), 806–913.
16. Kumar, P., Aruna, V., Pathamuthu, P., & Rajamani, K. (2025). An engineering framework for artificial intelligence-based marketing systems and digital consumer engagement models. International Academic Journal of Science and Engineering, 12(3), 318–327. https://doi.org/10.71086/IAJSE/V12I3/IAJSE1268
17. Mayuranathan, M., Akilandasowmya, G., Jayaram, B., Velrani, K. S., Kumar, M. J., & Vidhya, R. G. (2023). Artificial intelligent based models for event extraction using customer support applications. In 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) (pp. 167–172). IEEE.
18. Kotadiya, U., Arora, A. S., & Yachamaneni, T. (2021). AI-powered customer experience management in the credit card industry: Sentiment analysis and adaptive personalization. International Journal of Emerging Trends in Computer Science and Information Technology, 2(2), 35–44.
19. Ganesan, A., Velanganni, R., Rajapriya, M., Sundara Bala Murugan, P., & Paranthaman, P. (2025). The role of artificial intelligence in personalized marketing: Integrating AI with information systems for competitive

advantage. Indian Journal of Information Sources and Services, 15(1), 220–229. https://doi.org/10.51983/ijiss-2025.IJISS.15.1.28
20. Shaheen, A. (2025). Enhancing web user experience using AI-based personalization techniques. Journal of Natural and Applied Sciences, 1(10), 159–174.
21. Sethi, K., & Kapoor, M. (2024). Data-driven marketing in the age of AI: Reflections from the periodic series on technology and business integration. In Digital Marketing Innovations (pp. 7–11). Periodic Series in Multidisciplinary Studies.
22. Inavolu, S. M. (2024). Exploring AI-driven customer service: Evolution, architectures, opportunities, challenges and future directions. International Journal of Engineering and Advanced Technology, 13(3), 156–163.