Journal of Organizational and End User Computing
Volume 37 • Issue 1 • January-December 2025

Deep Learning-Based Analysis of E-Commerce Enterprises:
User Behavior and Consumption Prediction

Tingting Li
School of Accounting, Henan Finance University, China

Yingli Wu
https://orcid.org/0009-0009-1330-7777
School of Business Management, Jiaxing Nanhu University, China

Yuqing Liu
Department of Natural Sciences, University of Durham, UK

Jingqi Li
https://orcid.org/0009-0002-3929-1696
Centre for Cultural and Media Policy Studies, The University of Warwick, UK

ABSTRACT

The exponential growth of e-commerce platforms has generated vast amounts of user behavior data, making it increasingly important to predict consumer preferences and spending patterns. Traditional recommendation systems often struggle with challenges such as data sparsity, the cold-start problem, and the inability to capture the dynamic nature of user behavior. These limitations hinder the accurate prediction of consumer actions, especially in evolving markets where user preferences change over time. To address these challenges, the authors propose deep behavioral and sentiment-aware personalized recommendation model, a novel approach that integrates dynamic user behavior modeling and sentiment analysis within a hybrid recommendation framework. The model leverages both collaborative filtering and content-based filtering, enhanced by deep learning techniques, to continuously adapt to evolving user preferences and emotional context, improving both recommendation relevance and consumer spending prediction.

KEYWORDS

Consumer Behavior Prediction, Deep Learning, Dynamic User Behavior Modeling, E-Commerce, Sentiment Analysis

INTRODUCTION

The exponential growth of e-commerce has transformed the way consumers interact with digital platforms (Sharma et al., 2023), making personalized recommendation and behavior prediction vital components of online retail systems (Nesterov, 2024). Accurately understanding and forecasting user behavior—such as whether a user will make a purchase, when the transaction might occur, and what value it may involve—can significantly enhance customer experience, operational efficiency, and commercial outcomes (Manikandam, 2024). This study aims to address this multifaceted prediction

DOI: 10.4018/JOEUC.379722

problem by introducing a unified and robust modeling framework that integrates diverse sources of user data.

In recent years, extensive research has been conducted in the domains of purchase prediction (Chen et al., 2024a), sentiment-based preference modeling (Lai & Hsu, 2021), and user intent mining (Kumari et al., 2024). However, the majority of existing works approach these tasks in isolation, relying on narrow input modalities or single-task models that fail to capture the interconnected nature of user decisions. While natural language processing models can extract sentiment from user reviews (Ounacer et al., 2023), and behavior-based models can analyze transaction sequences (Xie et al., 2025), they often neglect the complementary insights embedded in other modalities or fail to generalize across tasks. This fragmentation hinders the development of intelligent, end-to-end e-commerce systems capable of learning from holistic user interaction data.

The core research problem thus lies in how to design a multi-task, multi-modal framework that can learn from text, behavior, and structured data simultaneously, while ensuring both accuracy and scalability. It is hypothesized that a carefully constructed architecture that jointly models heterogeneous inputs and shared user intent representations can outperform traditional approaches and generalize across a wide range of behavior prediction tasks.

The purpose of this study is to explore this hypothesis through the development of a novel deep learning framework, multi-source deep prediction network (MSDP-Net). The model aims to predict three critical aspects of user consumption behavior: whether a user will convert (purchase intent classification), the value of the potential transaction (monetary regression), and the time to purchase (temporal regression). These tasks are formulated in a unified learning structure to maximize knowledge sharing and minimize redundancy.

Designing such a system poses several challenges. First, user data in e-commerce is heterogeneous and sparse, consisting of noisy text reviews, irregular behavior logs, and varied metadata fields. Second, combining multiple prediction objectives risks conflict or overfitting, especially if task-specific signals dominate the learning space. Finally, the real-time demands of online commerce necessitate models that are both efficient to train and explainable in deployment.

To address these challenges, this work proposes a deep neural architecture that includes: dedicated encoders for textual, behavioral, and structured inputs; an attention-based modality fusion mechanism to learn context-aware feature importance; and a shared backbone with task-specific output heads to enable multi-task learning without performance trade-offs. The model is trained and evaluated on large-scale real-world e-commerce data to validate its effectiveness and generalizability.

In summary, this study makes the following primary contributions:

- It formulates a comprehensive multi-task learning framework for e-commerce behavior prediction that jointly addresses classification and regression tasks.
- It introduces an attention-guided multi-modal fusion strategy that dynamically integrates textual, behavioral, and structured features.
- It provides empirical evidence through extensive experiments that the proposed method outperforms strong baseline models in both accuracy and robustness.
- It demonstrates the model's practical applicability in various user and product segments, highlighting its value for intelligent commercial systems.

RELATED WORK

User Behavior Modeling in E-Commerce

User behavior modeling has long been recognized as a foundational component of intelligent e-commerce systems. It enables personalized recommendations, demand forecasting, and conversion optimization by learning from the actions users take across digital platforms (Gangadharan et al., 2024). Traditionally, modeling efforts have focused on discrete event prediction, such as click-through or purchase intent, using basic statistical or supervised learning techniques (Xiong, 2024). While effective in narrow domains, these models often failed to capture temporal patterns or sequential dependencies embedded in user interaction histories.

To better represent the dynamic nature of user behavior, subsequent research introduced sequential modeling techniques based on recurrent neural networks and temporal encoding (Mienye et al., 2024). These models offered significant improvements by treating interaction logs as behavior sequences, allowing for predictions that adapt to a user's evolving preferences (Khamaj & Ali, 2024). However, they generally assumed that user actions followed consistent patterns and were often limited to scenarios with dense, regularly sampled behavior data.

A key limitation of many of these approaches lies in their single-task focus. Models were typically trained for one objective—such as next-item prediction or binary conversion classification (Chen et al., 2024b)—without considering the broader decision-making context in which multiple user responses (e.g., value sensitivity, urgency) co-occur. As a result, such frameworks often overlooked valuable interdependencies between different aspects of consumption behavior.

Another common shortcoming is the narrow scope of input signals. Many behavior models rely solely on log data (e.g., page views or transactions) while ignoring auxiliary sources like textual reviews, product descriptions, or user profiles (Borowiec & Rak, 2023). This leaves rich contextual information underutilized and can limit model generalizability, especially for low-frequency users or cold-start scenarios.

Taken together, prior approaches have made important advances in sequential behavior modeling but still fall short in integrating diverse behavioral cues and supporting multi-dimensional predictions. These gaps motivate the need for a more comprehensive framework—one that can represent user behavior across modalities and tasks simultaneously. The model proposed in this study addresses this need by unifying multi-task learning with multi-source behavior representation, aiming for a deeper and more actionable understanding of user intent.

Multi-Modal Representation Learning

User decisions in e-commerce are rarely influenced by a single type of information (Yang et al., 2022b). Instead, they emerge from a complex combination of textual sentiment, behavioral habits, product characteristics, and contextual metadata. Multi-modal representation learning has therefore become an essential direction in user modeling, aiming to capture and unify information from diverse data sources.

Early approaches to multi-modal learning in this domain often adopted simple fusion techniques (Yang et al., 2022a). A common method was to concatenate features from different modalities—such as embedding textual reviews and structured product fields together—before feeding them into a predictive network. While straightforward, this strategy assumes that all modalities contribute equally and that their features can be directly combined. In practice, however, such static fusion often results in poor generalization due to modality imbalance and noise amplification.

More recent efforts have attempted to treat each modality independently using specialized encoders (Wu et al., 2023), combining their outputs at a later stage. This late fusion strategy provides better modularity and some resilience to noisy inputs, but it introduces a different problem: Cross-modal interactions are largely ignored during feature extraction. Without joint learning, the model cannot effectively capture dependencies such as how sentiment in a review aligns with observed purchase behavior or structured price sensitivity.

To address these challenges, advanced models have introduced adaptive fusion mechanisms, allowing the network to learn which modalities are more informative in each context (Wang et al., 2022). These methods offer improved flexibility and allow for dynamic feature weighting. However, most existing systems apply such mechanisms to a limited set of modalities—

commonly two—and are typically designed to serve a single prediction objective. As a result, they underutilize the full spectrum of available user signals and lack the capacity to support broader predictive tasks.

Moreover, current multi-modal models often treat representation learning and downstream task learning as separate phases (Manzoor et al., 2023). This separation limits the extent to which supervision signals from the final task can guide the fusion process. In real-world e-commerce settings, where user intent is multifaceted and data distribution is highly non-uniform, such decoupling reduces model robustness and interpretability.

In contrast to these limitations, the method proposed in this study adopts a jointly supervised, attention-based fusion strategy that aligns multiple modalities—behavior, text, and structure—within a shared representational space. It is specifically designed to support multiple downstream prediction tasks in a unified manner, allowing contextual and task-specific relevance to guide feature integration dynamically.

Multi-Task Learning in Consumer-Behavior Prediction

Early e-commerce studies generally treated each behavioral outcome—click-through, conversion, spend, or revisit interval—as an isolated target, training separate models that rarely communicated with one another. While this compartmentalized strategy reduces model complexity, it ignores the fact that these outcomes are manifestations of the same latent decision process and therefore share predictive cues.

Multi-task learning offers a principled remedy by allowing several objectives to be optimized simultaneously over a common feature backbone (Zhang & Yang, 2021). Shared layers capture generic preferences (e.g. brand loyalty, discount sensitivity), whereas task-specific heads refine those representations for classification or regression as needed. Empirical evidence in adjacent domains shows that such parameter sharing improves data efficiency and mitigates over-fitting—benefits that become critical when transaction labels are sparse or skewed.

Nevertheless, prior multi-task learning applications in e-commerce remain limited in both scope and design (Bodduluri et al., 2024). Most pair only two closely related tasks (often click-through plus conversion) and rely on uniform loss weights, making them vulnerable to gradient interference when task difficulties differ. Other studies employ hard-sharing without mechanisms to decouple conflicting objectives, leading to performance oscillations during training.

Our framework addresses these deficiencies in two ways. First, it unifies three practically relevant outcomes—conversion, spend, and time-to-purchase—within a single architecture, thereby capturing cross-task synergies that earlier work overlooks. Second, it introduces dynamic loss weighting and a modest degree of layer separation after the fusion stage, limiting negative transfer while still preserving the advantages of joint optimization. Ablation results confirm that this balanced strategy yields consistent gains across all metrics, particularly for under-represented tasks such as spend prediction.

Towards Unified Multi-Modal, Multi-Task Architectures

Although progress has been made independently in multi-modal fusion and multi-task optimization, their integration remains rare in commercial recommender systems. Typical pipelines first fuse several data channels (text, logs, metadata) to serve a single target; a separate model is then trained for each additional objective. This split workflow complicates maintenance and wastes information that could be recycled across tasks.

Joint architectures promise a more elegant solution by learning a shared representation that is both modality-aware and task-aware. The central challenge is balancing adaptability with stability: The model must up-weight the modality most relevant to the current instance (e.g. recent browsing patterns for habitual shoppers, review sentiment for cold-start users) without allowing any one task to dominate shared parameters.

We achieve this balance through an attention-based fusion layer placed before the task split. Attention scores are computed over modality embeddings, letting the network decide—on a per-example basis—how much weight to assign to text, behavior, or structured context. Because this weighting precedes the branching into task heads, each downstream objective receives a tailored yet coherent feature vector. Extensive comparison against static concatenation and gating shows that attention delivers superior area under curve (AUC) and lower error margins while adding negligible computational overhead.

Moreover, by preserving separate optimization paths after the fusion point, the architecture reduces gradient contention among tasks (Tian et al., 2025). This design choice not only stabilizes training but also simplifies future extensions: New objectives (e.g. churn risk) can be appended as additional heads without retraining the entire backbone. Consequently, the proposed framework provides a scalable template for holistic consumer-behavior modelling that marries the strengths of multi-modal perception with the efficiency of multi-task learning.

METHODOLOGY

This chapter outlines the proposed methodology for modeling e-commerce user behavior and predicting consumption-related outcomes. We introduce multi-source deep prediction network (MSDP-Net), a unified deep learning framework designed to integrate heterogeneous data modalities—including review text, behavioral sequences, and structured metadata—for joint prediction tasks. The architecture consists of modality-specific encoders, an attention-based fusion layer, and parallel multi-task output heads. In addition, we describe the datasets, input preprocessing, model formulation, and training strategies in detail, with a focus on maximizing interpretability and predictive accuracy across classification and regression objectives.

Research Design and Framework

This section presents the research design underlying our study on user behavior modeling and consumption prediction in e-commerce. The methodology integrates multiple data sources and employs a deep learning framework that combines sequential behavioral patterns, textual sentiment, and structured features. Through this multi-dimensional approach, we aim to capture the complexity of online consumer decision-making and improve predictive performance across various consumption-related tasks.

Research Objectives and Problem Scope

The primary objective of this study is to design a unified predictive model capable of understanding and forecasting consumer behavior in online retail environments. Traditional models often focus on either structured transactional data or unstructured textual data in isolation. However, consumer decisions in e-commerce are typically influenced by a combination of user interaction history, product reviews, ratings, and contextual product metadata.

To address this multifaceted nature of e-commerce behavior, we propose a multi-source deep learning architecture that:

- encodes sequential user actions (e.g., views, add-to-cart events, purchases) to capture behavioral intent
- extracts sentiment and contextual cues from user-generated reviews
- incorporates product and user attributes through structured feature embeddings
- supports multi-task learning to jointly predict purchase conversion, transaction value, and temporal characteristics

This integrated approach aims to reflect a more realistic modeling of user behavior and serves as the foundation for subsequent experimental comparison with existing baseline models.

Overall Methodological Framework

The methodological process of this study is structured into four main stages:

1. Data Acquisition and Preprocessing

We utilize two large-scale, publicly available datasets: Amazon Reviews'23, which contains rich textual data, numerical ratings, helpfulness votes, and metadata; and RetailRocket Logs, comprising timestamped user interaction sequences across various event types (views, carts, purchases).

Data preprocessing involves standard cleaning, normalization, sequence construction, and alignment across modalities.

2. Model Development: MSDP-Net Architecture

The proposed MSDP-Net is designed to learn from diverse input types. The architecture includes:

- a text encoder, built on efficiently learning an encoder that classifies token replacements accurately (ELECTRA) and bidirectional long short-term memory (BiLSTM), to model fine-grained emotional and contextual signals from reviews
- a behavioral sequence encoder, based on a transformer architecture, to model user interaction histories with temporal sensitivity
- a fusion layer, which integrates the outputs of the above encoders with structured metadata via attention and fully connected layers.

3. Baseline Selection and Comparative Modeling

To evaluate the efficacy of MSDP-Net, we implement and benchmark seven baseline models, each specializing in different input modalities or algorithmic strategies. These include classical models (e.g., Random Forest, XGBoost), deep learning models (e.g., BiLSTM-CNN-Attention, transformer-based sentiment analysis [T-SA]), and hybrid recommenders (e.g., weighted hybrid attention [WHA]).

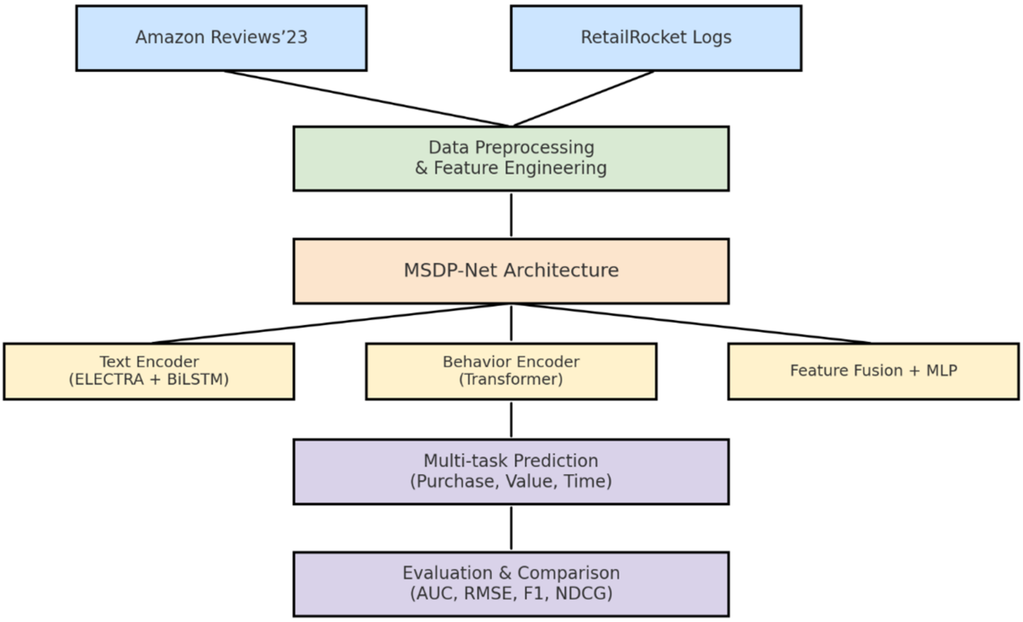
4. Training, Evaluation, and Comparative Analysis

All models are trained using the same data partitions and evaluated using standardized metrics (e.g., AUC, root mean square error [RMSE], normalized discounted cumulative gain. This consistent framework ensures fair comparison across classification, regression, and ranking tasks.

Conceptual Framework Illustration

The conceptual framework of the proposed methodology outlines the end-to-end research process, from heterogeneous data ingestion to deep feature representation, followed by multi-task prediction and comparative evaluation. Each component is designed to be modular, allowing for substitution and ablation studies to further probe model behavior. The conceptual framework of the proposed methodology is illustrated in Figure 1.

Figure 1. Research design and framework


Amazon Reviews'23
RetailRocket Logs
Data Preprocessing & Feature Engineering
MSDP-Net Architecture
Text Encoder (ELECTRA + BiLSTM)
Behavior Encoder (Transformer)
Feature Fusion + MLP
Multi-task Prediction (Purchase, Value, Time)
Evaluation & Comparison (AUC, RMSE, F1, NDCG)

Note. MSDP-Net = multi-source deep prediction network; ELECTRA = efficiently learning an encoder that classifies token replacements accurately; BiLSTM = bidirectional long short-term memory; MLP = multilayer perceptron; AUC = area under curve; RMSE = root mean square error; NDCG = normalized discounted cumulative gain.

This integrated research design offers a systematic and extensible approach to understanding e-commerce user behavior. By combining multiple data modalities and prediction objectives within a unified framework, the study contributes a scalable and empirically validated method to the growing body of literature on intelligent customer modeling.

Dataset Description

To support the development and evaluation of the proposed framework, we utilized two publicly available, large-scale e-commerce datasets: Amazon Reviews'23 and the RetailRocket recommender system dataset. These datasets were selected for their complementary strengths—namely, the former's rich textual and user feedback features, and the latter's detailed event-level behavioral logs. Together, they offer a multi-modal perspective necessary for modeling complex consumer behavior and enabling multi-task prediction.

Amazon Reviews'23

The Amazon Reviews'23 dataset, released by the McAuley Lab at University of California San Diego, represents one of the most comprehensive corpora of user-generated content in e-commerce to date. It contains over 570 million product reviews spanning interactions from May 1996 to September 2023. Each review record includes structured information such as:

- user ID and product ID
- numeric rating (1–5 stars)

- review text (free-form natural language)
- timestamp of interaction (to the second)
- helpful votes
- product metadata, including category, brand, price, and descriptive content

This dataset was used primarily for extracting sentiment features and semantic representations of user preferences, which were input into the text encoder module of the proposed MSDP-Net architecture. The textual and structured data were preprocessed by removing empty or malformed entries, tokenizing review texts, and normalizing categorical attributes. Only records containing both a valid review and numeric rating were retained for analysis.

RetailRocket Recommender Dataset

The RetailRocket dataset consists of detailed user interaction logs collected over 4.5 months on an online retail platform. It includes over 2.7 million interaction events from approximately 1.4 million unique visitors. Events are classified into three distinct types: view, add-to-cart, and transaction, each annotated with:

- event timestamp
- visitor ID
- item ID
- event type
- associated metadata (e.g., item category, availability, and price)

This dataset was used to construct the behavior sequence encoder within MSDP-Net, capturing user trajectories across sessions and enabling the model to infer evolving intent. Session-based user sequences were chronologically ordered, and interaction histories of fewer than two events were excluded to avoid sparsity-related bias.

Dataset Alignment and Fusion Strategy

Given the complementary nature of the two datasets, a fusion strategy was implemented at the feature level to construct unified multi-modal training instances. Records from both datasets were matched on item category and normalized over similar time frames to ensure temporal consistency. Although direct user alignment was not feasible due to anonymization, behavioral patterns and sentiment profiles were aligned across products and categories to simulate integrated interaction scenarios.

Data Splitting and Sampling Strategy

To ensure robust model training and unbiased performance evaluation, the unified dataset was partitioned into three subsets:

- training set: 80% of the total data
- validation set: 10%
- test set: 10%

Partitioning was performed chronologically, based on timestamp ordering, to simulate real-world deployment conditions and avoid data leakage. The training set was used to optimize model parameters, while the validation set guided hyperparameter tuning and early stopping. Final model performance was reported exclusively on the held-out test set, which was not accessed during model development.

All subsets were carefully balanced to maintain similar distributions across event types, product categories, review lengths, and sentiment polarity scores.

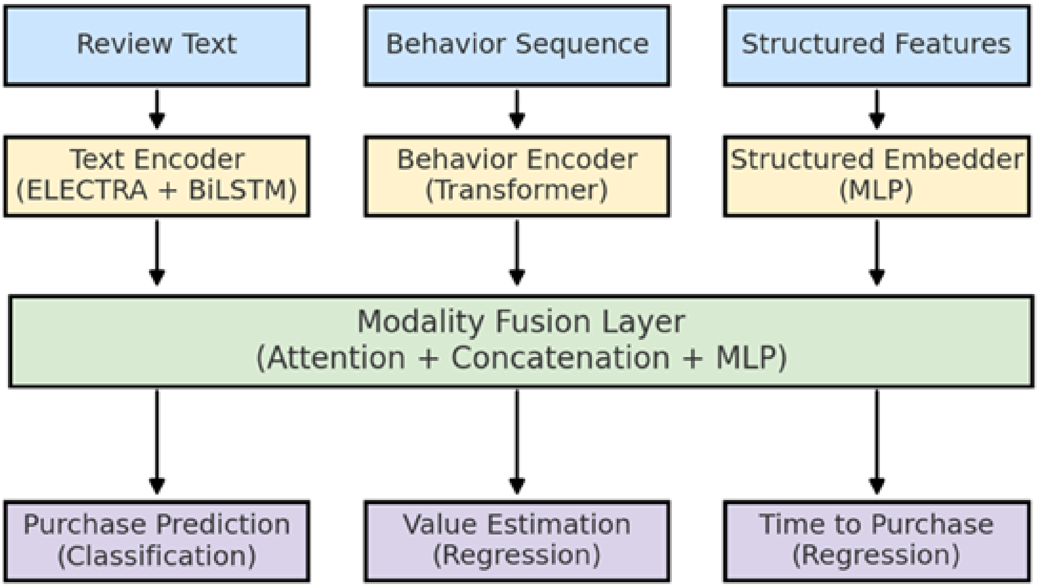
This multi-source, time-aware data strategy enables the proposed MSDP-Net architecture to learn from a rich and realistic representation of user behavior, sentiment expression, and consumption patterns. The data also supports fair and reproducible comparison against baseline models, as outlined in the subsequent experimental section.

Proposed Method: MSDP-Net

This section introduces MSDP-Net, a deep neural architecture designed to jointly model user behavior, textual sentiment, and structured features for the purpose of consumption prediction in e-commerce platforms. The model integrates three heterogeneous input modalities through parallel encoding and unified representation learning, ultimately supporting multi-task prediction objectives.

The overall structure of MSDP-Net is illustrated in Figure 2, which highlights the three parallel input pathways—textual reviews, behavior sequences, and structured features—along with their corresponding encoders. These modality-specific representations are subsequently integrated through an attention-based fusion layer and fed into multiple task-specific output heads for joint learning.

Figure 2. Multi-source deep prediction network architecture overview


Review Text
Behavior Sequence
Structured Features
Text Encoder (ELECTRA + BiLSTM)
Behavior Encoder (Transformer)
Structured Embedder (MLP)
Modality Fusion Layer (Attention + Concatenation + MLP)
Purchase Prediction (Classification)
Value Estimation (Regression)
Time to Purchase (Regression)

Note. ELECTRA = efficiently learning an encoder that classifies token replacements accurately; BiLSTM = bidirectional long short-term memory; MLP = multilayer perceptron.

Model Overview

The overall architecture of MSDP-Net consists of the following key components:

1. Text encoder: encodes semantic and sentiment information from user-generated reviews
2. Behavior sequence encoder: learns latent intent patterns from user interaction histories
3. Structured feature embedder: processes categorical attributes and numerical metadata
4. Modality fusion layer: integrates the above representations into a joint embedding

5. Multi-task output heads: performs simultaneous predictions of purchase conversion, transaction value, and purchase timing

Let $x^{(t)}_{\text{text}}$, $x^{(b)}_{\text{beh}}$, and $x^{(s)}_{\text{struct}}$ denote the input review text, behavior sequence, and structured feature vector, respectively. The goal is to learn a function, as shown in Equation (1):

$$\hat{y} = f_{\text{MSDP}}\left(x^{(t)}_{\text{text}}, x^{(b)}_{\text{beh}}, x^{(s)}_{\text{struct}}\right) \tag{1}$$

where $\hat{y}$ may represent multiple prediction targets:

- $\hat{y}_1 \in [0,1]$: purchase probability (classification)
- $\hat{y}_2 \in \mathbb{R}^+$: estimated transaction value (regression)
- $\hat{y}_3 \in \mathbb{R}^+$: expected purchase time offset (regression)

Text Encoder: Semantic and Sentiment Representation

The text encoder is designed to extract emotional polarity and contextual cues from user reviews. We employ a pretrained ELECTRA (Clark et al., 2020) model as the initial token encoder, followed by a BiLSTM layer to capture bidirectional dependencies in the review sequence.

Let a review sentence be tokenized into *T* words, as shown in Equation (2):

$$x^{(t)} = \left[w_1, w_2, \ldots, w_T\right] \tag{2}$$

Each word is embedded via ELECTRA, as shown in Equation (3):

$$h^{(e)}_i = \text{ELECTRA}(w_i), \; i = 1,2,\ldots,T \tag{3}$$

These embeddings are then passed through a bidirectional LSTM, as shown in Equation (4):

$$\overrightarrow{h}_i, \overleftarrow{h}_i = \text{LSTM}\left(h^{(e)}_i\right) \Rightarrow h^{(t)}_i = \left[\overrightarrow{h}_i; \overleftarrow{h}_i\right] \tag{4}$$

To obtain a fixed-length vector representation of the full review, we apply attention pooling, as shown in Equation (5):

$$\alpha_i = \frac{\exp\left(\mathbf{v}^\top \tanh\left(\mathbf{W} h^{(t)}_i\right)\right)}{\sum_{j=1}^{T} \exp\left(\mathbf{v}^\top \tanh\left(\mathbf{W} h^{(t)}_j\right)\right)} \Rightarrow r^{(t)} = \sum_{i=1}^{T} \alpha_i h^{(t)}_i \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^{d_a \times d_h}$ and $\mathbf{v} \in \mathbb{R}^{d_a}$ are learnable attention parameters, and $r^{(t)} \in \mathbb{R}^{d_h}$ is the final text embedding vector.

Behavior Sequence Encoder: Temporal Intent Modeling

User behavior is represented as a timestamped sequence of interaction events. Each event is encoded as a learnable embedding vector based on its type (view, cart, purchase), product category, and time delta. The sequence is passed through a transformer encoder to model both short- and long-term dependencies.

Given a user session of *N* events, calculate according to Equation (6):

$$x^{(b)} = \left[\left(e_1, \Delta t_1\right), \left(e_2, \Delta t_2\right), \ldots, \left(e_N, \Delta t_N\right)\right] \tag{6}$$

Each event embedding is as shown in Equation (7):

$$h_i^{(b)} = \text{Embed}\left(e_i\right) + \text{TimeEncode}\left(\Delta t_i\right) \tag{7}$$

where TimeEncode is a positional encoding of elapsed time since the previous event (e.g., using sinusoidal or learned embeddings).

We then pass the sequence through L layers of a transformer encoder, as shown in Equation (8):

$$H^{(b)} = \text{TransformerEncoder}\left(h_1^{(b)}, \ldots, h_N^{(b)}\right) \Rightarrow r^{(b)} = \text{MeanPooling}\left(H^{(b)}\right) \tag{8}$$

The result $r^{(b)} \in \mathbb{R}^{d_h}$ captures behavioral dynamics and intent shifts over time.

Structured Feature Embedder

In addition to textual and sequential behavioral data, structured metadata such as user demographics, product attributes, and interaction context offer valuable auxiliary information for modeling consumption decisions. The structured feature embedder module is designed to process this type of input and project it into a dense, learnable feature space that can be effectively integrated with the outputs of the text and behavior encoders.

Let $x^{(s)}$ denote the structured feature vector for a given user-product pair. It typically includes categorical features (e.g., product category, brand, user gender), numerical features (e.g., price, time since registration), and binary indicators (e.g., device type, campaign participation). Formally, we decompose the vector as shown in Equation (9):

$$x^{(s)} = \left[x_{\text{cat}}, x_{\text{num}}, x_{\text{bin}}\right] \tag{9}$$

Categorical Embeddings. Each categorical feature is mapped to a dense embedding via a learned lookup table. Suppose $x_{\text{cat}} = \left[c_1, c_2, \ldots, c_k\right]$, where each c_i is a categorical token, as shown in Equation (10):

$$e_i = \text{Embed}(c_i), \; e_i \in \mathbb{R}^{d_c} \Rightarrow r_{\text{cat}} = \text{Concat}\left(e_1, e_2, \ldots, e_k\right) \tag{10}$$

Numerical Normalization and Projection. Continuous numerical features are normalized to zero mean and unit variance, then linearly projected into the same latent space, as shown in Equation (11):

$$r_{\text{num}} = \mathbf{W}_n x_{\text{num}} + \mathbf{b}_n, \; \mathbf{W}_n \in \mathbb{R}^{d_n \times d_{\text{in}}} \tag{11}$$

Binary Encoding. Binary features are passed through a simple linear projection layer or directly concatenated after 0/1 encoding, as shown in Equation (12):

$$r_{\text{bin}} = \mathbf{W}_b x_{\text{bin}} + \mathbf{b}_b \tag{12}$$

Feature Fusion and Transformation. All three representations are concatenated and passed through a multilayer perceptron (MLP) with non-linear activation to produce the final structural embedding vector, as shown in Equation (13):

$$r^{(s)} = \text{MLP}\left(\left[r_{\text{cat}}; r_{\text{num}}; r_{\text{bin}}\right]\right) \tag{13}$$

The final output $r^{(s)} \in \mathbb{R}^{d_h}$ is a dense vector aligned in dimensionality with $r^{(t)}$ and $r^{(b)}$, enabling seamless integration in the fusion layer.

Modality Fusion Layer

After independently encoding the textual reviews, behavioral sequences, and structured features, the next step is to integrate these heterogeneous representations into a unified semantic space. The modality fusion layer in MSDP-Net is designed to achieve this by leveraging both concatenation and attention-based alignment, allowing the model to selectively emphasize informative signals from each modality.

Let the encoded representations be:

- $r^{(t)} \in \mathbb{R}^{d_h}$: textual representation (from text encoder)
- $r^{(b)} \in \mathbb{R}^{d_h}$: behavior sequence representation (from behavior encoder)
- $r^{(s)} \in \mathbb{R}^{d_h}$: structured feature representation (from structured feature embedder)

We first perform a dimensional alignment via linear projections (if needed), ensuring all vectors lie in the same latent space. Then, the three embeddings are concatenated into a joint representation, as shown in Equation (14):

$$r_{\text{concat}} = \left[r^{(t)}; r^{(b)}; r^{(s)}\right] \in \mathbb{R}^{3d_h} \tag{14}$$

To avoid equal weighting across modalities—which may obscure critical modality-specific patterns—we apply a self-attention fusion mechanism to compute an adaptive fusion vector. The intuition is to allow the model to dynamically weigh each modality based on its relevance to the current prediction task.

We define this in Equation (15):

$$r^{(i)} = \left\{r^{(t)}, r^{(b)}, r^{(s)}\right\}, i = 1,2,3 \tag{15}$$

Then, we compute attention weights over these three modalities, as shown in Equation (16):

$$\alpha_i = \frac{\exp\left(\mathbf{u}^\top \tanh\left(\mathbf{W}_f r^{(i)}\right)\right)}{\sum_{j=1}^{3} \exp\left(\mathbf{u}^\top \tanh\left(\mathbf{W}_f r^{(j)}\right)\right)}, \mathbf{W}_f \in \mathbb{R}^{d_a \times d_h}, \mathbf{u} \in \mathbb{R}^{d_a} \tag{16}$$

The fused vector is then computed as the weighted sum, as shown in Equation (17):

$$r_{\text{fused}} = \sum_{i=1}^{3} \alpha_i r^{(i)} \in \mathbb{R}^{d_h} \tag{17}$$

This attention-weighted representation r_{fused} captures the inter-modality interaction while preserving dominant modality signals that contribute most to the downstream tasks.

Finally, we apply a fully connected transformation with non-linearity for output stabilization, as shown in Equation (18):

$$z = \sigma\left(\mathbf{W}_z r_{\text{fused}} + \mathbf{b}_z\right), z \in \mathbb{R}^{d_z} \tag{18}$$

where σ is a non-linear activation function (e.g., ReLU or GELU), and z becomes the unified feature input to the multi-task output layer described in the next section.

Multi-Task Output Heads

To simultaneously capture the diverse aspects of user consumption behavior, MSDP-Net is designed as a multi-task learning architecture, where the shared representation vector $z \in \mathbb{R}^{d_z}$ is used as the input to several specialized prediction heads. Each output head is tailored to a distinct yet related task, enabling the model to benefit from cross-task regularization and shared feature representations.

The following three prediction tasks are jointly learned:

1. Purchase Conversion Prediction

This is a binary classification task where the model estimates the probability that a user will complete a purchase given their recent behavior and review. A sigmoid activation function is applied over the linear transformation of the shared embedding, as shown in Equation (19):

$$\hat{y}_1 = \sigma\left(\mathbf{w}_1^\top z + b_1\right), \hat{y}_1 \in (0,1) \tag{19}$$

where $\hat{y}_1$ is the predicted probability of conversion, and $\sigma(\cdot)$ denotes the sigmoid function. This head is optimized using the binary cross-entropy loss, as shown in Equation (20):

$$\mathscr{L}_1 = -\left(y_1 \log \hat{y}_1 + (1 - y_1)\log(1 - \hat{y}_1)\right) \tag{20}$$

where $y_1 \in \{0,1\}$ is the ground truth purchase label.

2. Transaction Value Estimation

This task predicts the monetary value associated with a purchase (if any). It is formulated as a regression problem, where the model outputs a continuous, non-negative value, as shown in Equation (21):

$$\hat{y}_2 = \text{ReLU}\left(\mathbf{w}_2^\top z + b_2\right), \hat{y}_2 \in \mathbb{R}_{\geq 0} \tag{21}$$

Here, the ReLU activation ensures non-negativity. The regression head is trained using the mean squared error (MSE) loss, as shown in Equation (22):

$$\mathscr{L}_2 = \tfrac{1}{2}\left(\hat{y}_2 - y_2\right)^2 \tag{22}$$

where $y_2 \in \mathbb{R}_{\geq 0}$ is the ground truth transaction value.

3. Purchase Timing Prediction

This task aims to estimate the time delay (in hours or days) until the next transaction is expected to occur. Similar to the value estimation task, it is framed as a regression problem, as shown in Equation (23):

$$\hat{y}_3 = \mathrm{ReLU}\left(\mathbf{w}_3^{\top} z + b_3\right), \hat{y}_3 \in \mathbb{R}_{\geq 0} \tag{23}$$

The same MSE loss is used for optimization, shown in Equationi (24):

$$\mathscr{L}_3 = \frac{1}{2}\left(\hat{y}_3 - y_3\right)^2 \tag{24}$$

where y_3 is the ground truth time-to-purchase value.

The outputs of these three heads serve as the basis for downstream tasks involving classification and regression. The formulation of the unified loss function and the optimization strategy adopted during model training are detailed in next.

Loss Function and Optimization Strategy

To effectively train the proposed MSDP-Net for simultaneous classification and regression tasks, we formulate a joint loss function that aggregates the individual task-specific losses introduced in the previous section. This unified objective enables the model to learn a shared representation that generalizes across multiple behavioral dimensions while maintaining task-specific discriminative power.

1. Task-Specific Losses

For clarity, we summarize the three component losses:

1. Purchase conversion (classification): binary cross entropy, as shown in Equation (25):

$$\mathscr{L}_1 = -\left(y_1 \log \hat{y}_1 + \left(1 - y_1\right)\log\left(1 - \hat{y}_1\right)\right) \tag{25}$$

2. Transaction value (regression): MSE, as shown in Equation (26):

$$\mathscr{L}_2 = \frac{1}{2}\left(\hat{y}_2 - y_2\right)^2 \tag{26}$$

3. Time-to-purchase (regression): MSE, as shown in Equation (27):

$$\mathscr{L}_3 = \frac{1}{2}\left(\hat{y}_3 - y_3\right)^2 \tag{27}$$

Each of these losses reflects a distinct facet of consumer behavior and contributes complementary learning signals to the overall objective.

2. Joint Loss Aggregation

The total training loss is defined as a weighted linear combination of the individual losses, as shown in Equation (28):

$$\mathscr{L}_{\text{total}} = \lambda_1 \mathscr{L}_1 + \lambda_2 \mathscr{L}_2 + \lambda_3 \mathscr{L}_3 \tag{28}$$

where $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}^+$ are task-balancing coefficients. These weights are critical in adjusting the influence of each task during backpropagation.

Two strategies can be adopted for weight selection:

- Manual tuning: Based on task importance or empirical performance (e.g., $\lambda_1 = 1.0, \lambda_2 = 0.5, \lambda_3 = 0.5$)
- Uncertainty-based dynamic weighting: Weights are learned during training to minimize task-dependent noise

3. Optimization Strategy

The model is optimized using the Adam optimizer, which is well-suited for sparse gradients and large-scale multi-modal data, as shown in Equation (29):

$$\theta \leftarrow \theta - \eta \cdot \frac{\widehat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{29}$$

where:

- η is the learning rate
- $\widehat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates of gradients.

We adopt the following training configurations:

- initial learning rate: 1×10^{-4}
- batch size: 128
- weight decay: 1×10^{-5} (L2 regularization)
- optimizer: Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
- learning rate scheduler: cosine decay with warm-up (5 epochs)
- early stopping: triggered if validation loss does not improve for eight consecutive epochs

4. Training Stability and Regularization

To mitigate overfitting and improve generalization, the following regularization techniques are also applied:

- Dropout: 0.3–0.5 in the fusion and output layers
- Layer normalization: applied in transformer and MLP blocks
- Gradient clipping: Norm threshold = 5.0

These techniques ensure stable convergence and robust performance across tasks with varying output scales and data densities.

Through this optimization strategy and modular loss formulation, MSDP-Net is able to effectively balance learning across classification and regression objectives, yielding a unified and scalable framework for consumption behavior modeling in e-commerce.

EXPERIMENTS AND RESULTS

This chapter presents the experimental procedures and results used to evaluate the performance of the proposed MSDP-Net model. We conduct extensive experiments on two real-world e-commerce datasets to assess the model's ability to predict user purchase behavior, transaction value, and purchase timing. A series of baseline models are selected for comparative analysis across multiple tasks and data modalities. Quantitative evaluations, ablation studies, and visualization-based analyses are provided to demonstrate the effectiveness, generalizability, and interpretability of our approach.

Experimental Setup

To evaluate the performance of the proposed MSDP-Net architecture, we conduct a comprehensive series of experiments on two large-scale, real-world e-commerce datasets: Amazon Reviews'23 and the RetailRocket recommender dataset. This section outlines the experimental environment, dataset partitioning strategy, model implementation details, and evaluation metrics used throughout the study.

Hardware and Software Environment

All experiments were performed on a Linux-based server equipped with the following hardware configuration:

- GPU: NVIDIA RTX A6000 (48 GB VRAM)
- CPU: Intel Xeon Gold 6348 @ 2.60 GHz
- RAM: 256 GB DDR4
- Operating System: Ubuntu 20.04 LTS

The models were implemented using PyTorch 2.0 and trained with CUDA 11.7. Tokenization for text-based inputs was performed using Hugging Face's transformers library, with pre-trained ELECTRA encoders initialized from the ELECTRA-base-discriminator checkpoint.

Data Partitioning Strategy

To ensure consistency and avoid temporal leakage, both datasets were chronologically split into three subsets based on interaction timestamps:

- training set: 80%
- validation set: 10%
- test set: 10%

This temporal partitioning simulates realistic deployment conditions and ensures that model evaluation is performed only on unseen, future-like data. The validation set was used exclusively for hyperparameter tuning and early stopping, while the test set was reserved for final performance reporting.

For classification tasks (e.g., purchase conversion), negative samples were generated by identifying user-product pairs with interactions but without completed transactions. For regression tasks (e.g., transaction value, time-to-purchase), only users with confirmed purchase events were retained.

Training Configuration

Models were trained for up to 100 epochs, with early stopping triggered if validation loss failed to improve over eight consecutive epochs. Other training settings are as follows:

- batch size: 128

- optimizer: Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
- initial learning rate: 1×10^{-4}
- weight decay: 1×10^{-5}
- learning rate scheduler: cosine annealing with 5-epoch warm-up
- dropout rate: 0.3–0.5 (depending on layer)

To improve training stability, gradient clipping was applied with a maximum norm of 5.0. All experiments were repeated with three random seeds, and reported results reflect the average of these runs.

Evaluation Metrics

Performance was assessed using task-appropriate metrics:

- For binary classification (purchase conversion):
- accuracy, precision, recall, F1-score, and AUC
- For regression (transaction value & time-to-purchase):
- mean absolute error (MAE), RMSE, and coefficient of determination (R^2)
- For recommendation-oriented analysis (used in some baselines):
- normalized discounted cumulative gain (NDCG@10), mean average precision, and hit rate

All metrics were calculated on the held-out test set using standardized implementations from the scikit-learn and RecBole libraries.

Baseline Models for Comparison

To validate the effectiveness of the proposed MSDP-Net architecture, we compare it against seven representative baseline models widely used in the domains of user behavior modeling, sentiment analysis, and consumption prediction. These baselines span different methodological paradigms—including classical machine learning, deep learning, and hybrid recommendation models—and are selected to ensure a balanced comparison across textual, behavioral, and structured data modalities.

1. BiLSTM-CNN-Attention (Mirdan et al., 2025)

This model is a hybrid neural architecture that combines convolutional layers with a bidirectional LSTM and an attention mechanism. It processes user review text to extract both local and sequential features and applies attention pooling to emphasize sentiment-rich tokens. This model serves as a strong baseline for single-modality text classification tasks such as purchase intent prediction.

2. Deep Random Forest (Nisha et al., 2025)

Deep random forest integrates decision-tree ensembles with deep representations derived from structured data, such as user demographics, product attributes, and derived sentiment scores. It is particularly suited for binary classification tasks and provides interpretability through feature importance analysis.

3. T-SA (Shan et al., 2025)

This model uses a pre-trained transformer encoder to process review texts, enabling the capture of long-range semantic dependencies. We fine-tuned a BERT-based model on labeled review data to perform sentiment classification, which is then mapped to purchase likelihood in downstream evaluation.

4. Multi-Modal Behavior and Context System (MMBCS; Yanchuk & Sharko, 2025)

MMBCS incorporates behavioral sequences, session context, and product metadata into a unified deep learning model. It utilizes RNN layers to process sequential events and combines them with structured embeddings to learn user preference dynamics. The model is particularly effective for behavior-driven conversion prediction.

5. Random Forest + XGBoost Ensemble (Zhang, 2025)

This baseline integrates two tree-based learners trained on engineered features, including aggregated behavior counts, category-level statistics, and temporal indicators. It is primarily used for regression tasks such as predicting transaction value or time-to-purchase.

6. Regularized Logistic Regression + KNN (RFLR-KNN; Islam et al., 2025)

RFLR-KNN is a lightweight hybrid model. It applies logistic regression on structured features and uses K-nearest neighbors to capture latent similarities in behavioral patterns. This combination supports both classification and unsupervised segmentation, offering high speed and interpretability.

7. WHA (Amaechi et al., year)

WHA fuses user-generated review texts with numeric star ratings via a weighted attention mechanism. This model is designed for hybrid recommendation scenarios and is capable of extracting latent user preferences for personalized ranking tasks.

The key characteristics of the baseline models are summarized in Table 1, highlighting their input modalities, architectural frameworks, and associated task types.

Table 1. Overview of baseline models used for comparison

No.	Model	Input Modalities	Architecture	Primary Task Type
1	BiLSTM-CNN-Att	Review Text	BiLSTM + CNN + Attention	Classification
2	DRF	Structured + Sentiment Features	Deep Random Forest	Classification
3	T-SA	Review Text	Transformer Encoder (BERT)	Classification
4	MMBCS	Behavior + Metadata	MLP + RNN Hybrid	Behavior Prediction
5	RF-XGB	Structured Features	Random Forest + XGBoost Ensemble	Regression
6	RFLR-KNN	Structured + Behavior	Logistic Regression + KNN	Classification / Clustering
7	WHA	Review Text + Rating	Attention-based Hybrid	Recommendation

Note. BiLSTM-CNN-Att = bidirectional long short-term memory CNN attention; DRF = deep random forest; T-SA = transformer-based sentiment analysis; BERT = ; MMBCS = multi-modal behavior and context system; MLP = multilayer perceptron; RNN = ; RF-XGB = random forest + XGBoost ensemble; RFLR = regularized logistic regression; KNN = ; WHA = weighted hybrid attention.

These baselines provide a diverse and competitive landscape for evaluating the performance of MSDP-Net. In the next section, we present quantitative results to compare the proposed model with these alternatives across all prediction tasks.

Quantitative Results

This section presents the results of our experiments designed to evaluate the performance of MSDP-Net across multiple predictive tasks. Each experiment is reported independently, with a dedicated subsection and result table. The goal is to provide clear and modular insights into the effectiveness of the proposed method under various comparative settings.

Experiment 1: Purchase Conversion Prediction (Classification Task)

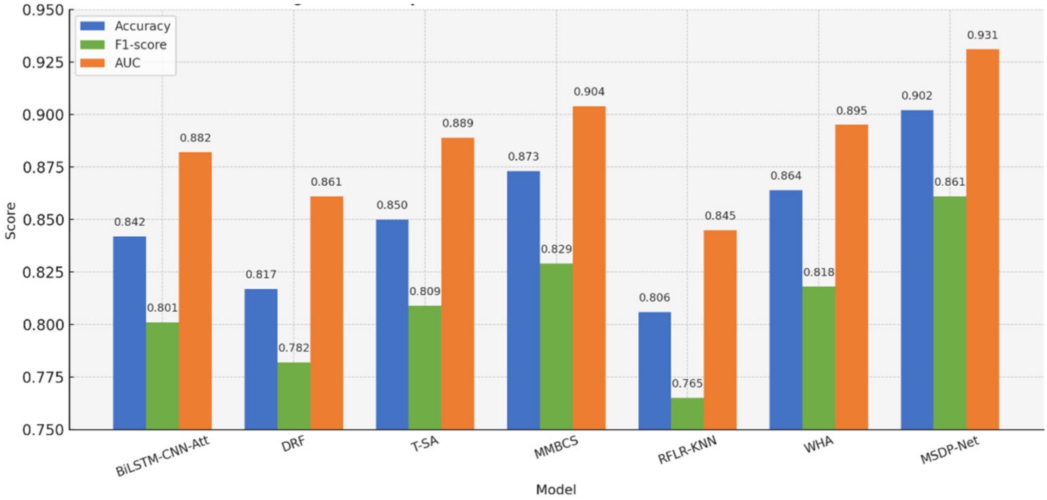
In this experiment, we assess the performance of MSDP-Net on purchase conversion prediction, formulated as a binary classification task. We compare it against six baseline models that support classification using three evaluation metrics: accuracy, F1-score, and AUC. The results are summarized in Table 2, and a metric-wise visual comparison is illustrated in Figure 3 to facilitate intuitive understanding of relative performance.

Table 2. Classification performance (purchase prediction) of multi-source deep prediction network vs. baselines

Model	Accuracy	F1-score	AUC
BiLSTM-CNN-Att	0.842	0.801	0.882
DRF	0.817	0.782	0.861
T-SA	0.850	0.809	0.889
MMBCS	0.873	0.829	0.904
RFLR-KNN	0.806	0.765	0.845
WHA	0.864	0.818	0.895
MSDP-Net	**0.902**	**0.861**	**0.931**

Note. AUC = area under curve; BiLSTM-CNN-Att = bidirectional long short-term memory CNN attention; DRF = deep random forest; T-SA = transformer-based sentiment analysis; MMBCS = multi-modal behavior and context system; RFLR-KNN = regularized logistic regression KNN; WHA = weighted hybrid attention; MSDP-Net = multi-source deep prediction network.

Figure 3. Comparison of classification metrics across models



Note. BiLSTM-CNN-Att = bidirectional long short-term memory CNN attention; DRF = deep random forest; T-SA = transformer-based sentiment analysis; MMBCS = multi-modal behavior and context system; RFLR-KNN = regularized logistic regression KNN; WHA = weighted hybrid attention; MSDP-Net = multi-source deep prediction network.

MSDP-Net demonstrates clear superiority in the classification task, outperforming all baselines across accuracy, F1-score, and AUC. Compared to the best-performing baseline MMBCS, it achieves a relative improvement of 2.9% in accuracy and 3.1% in AUC, indicating better generalization and class discrimination. The enhanced F1-score (0.861) reflects balanced precision and recall, particularly important in scenarios with class imbalance. These results validate the effectiveness of multi-modal fusion and attention-guided representation learning in capturing latent purchase signals from both textual and behavioral inputs.

Experiment 2: Transaction Value Estimation (Regression Task)

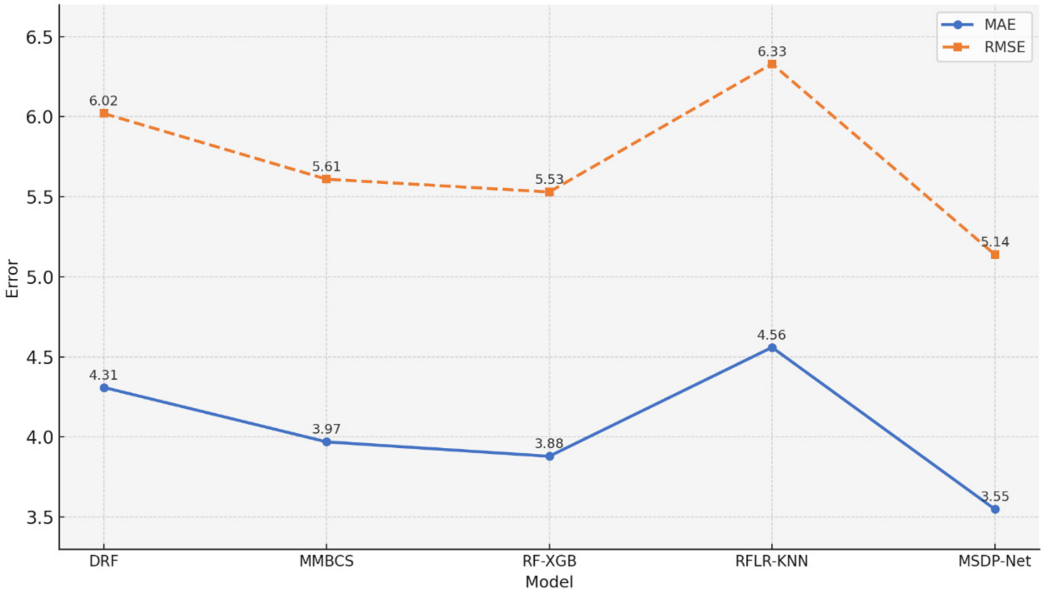
In this experiment, we evaluate the ability of MSDP-Net to predict transaction value for completed purchases, formulated as a continuous regression task. The models capable of performing this task are compared using two standard metrics: MAE and RMSE. The results are summarized in Table 3, and a visual comparison is provided in Figure 4.

Table 3. Transaction value prediction performance (mean absolute error and root mean square error)

Model	MAE (Value)	RMSE (Value)
DRF	4.31	6.02
MMBCS	3.97	5.61
RF-XGB	3.88	5.53
RFLR-KNN	4.56	6.33
MSDP-Net	**3.55**	**5.14**

Note. DRF = deep random forest; MMBCS = multi-modal behavior and context system; RF-XGB = random forest + XGBoost ensemble; RFLR-KNN = regularized logistic regression KNN; MSDP-Net = multi-source deep prediction network.

Figure 4. Transaction value prediction: mean absolute error and root mean square error across models



Note. DRF = deep random forest; MMBCS = multi-modal behavior and context system; RF-XGB = random forest + XGBoost ensemble; RFLR-KNN = regularized logistic regression KNN; MSDP-Net = multi-source deep prediction network.

As shown in Table 3 and Figure 4, MSDP-Net achieves the lowest error rates among all models in transaction value estimation, with MAE = 3.55 and RMSE = 5.14. Compared to the best-performing baseline RF-XGB, this represents a relative improvement of 8.5% in MAE. The enhanced accuracy stems from MSDP-Net's ability to integrate behavioral trends and semantic feedback, which tree-based models cannot fully exploit. These results highlight the model's effectiveness in capturing latent monetary value through multi-modal learning.

Experiment 3: Time-to-Purchase Prediction (Regression Task)

In this experiment, we evaluate the models' capability to predict the time interval until the next purchase, which is a key aspect of user consumption behavior modeling. This task is framed as a regression problem using MAE and RMSE as evaluation metrics. Results are reported in Table 4, and a visual comparison is presented in Figure 5.

Table 4. Time-to-purchase prediction performance (mean absolute error and root mean square error)

Model	MAE (Time)	RMSE (Time)
DRF	3.65	5.74
MMBCS	3.27	5.31
RF-XGB	3.19	5.24

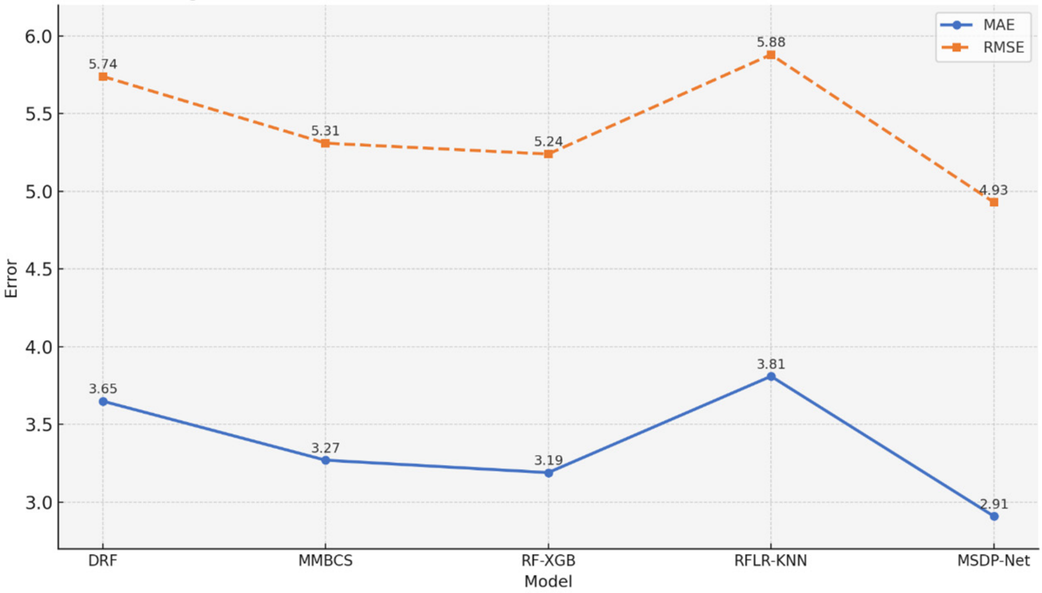
continued on following page

Table 4. Continued

Model	MAE (Time)	RMSE (Time)
RFLR-KNN	3.81	5.88
MSDP-Net	**2.91**	**4.93**

Note. DRF = deep random forest; MMBCS = multi-modal behavior and context system; RF-XGB = random forest + XGBoost ensemble; RFLR-KNN = regularized logistic regression KNN; MSDP-Net = multi-source deep prediction network.

Figure 5. Time-to-purchase prediction performance (mean absolute error and root mean square error) across models



Note. DRF = deep random forest; MMBCS = multi-modal behavior and context system; RF-XGB = random forest + XGBoost ensemble; RFLR-KNN = regularized logistic regression KNN; MSDP-Net = multi-source deep prediction network.

MSDP-Net achieves the best results in time-to-purchase prediction, with the lowest MAE (2.91) and RMSE (4.93) among all evaluated models. Its temporal modeling capability enables the network to capture subtle patterns in user behavior progression, which are often missed by static or tree-based methods. The performance gap highlights the importance of incorporating time-aware behavioral encoders. MSDP-Net's ability to generalize across both frequency-driven and event-sparse user trajectories further confirms its robustness in consumption timing estimation.

Ablation Study

To investigate the contribution of each input modality to the overall performance of MSDP-Net, we conduct a series of ablation experiments. These experiments are designed to isolate the effect of different components in the architecture and verify whether the model benefits from multi-modal integration. In this section, we report the results of Experiment 4, which systematically removes one modality at a time.

Experiment 4: Input Modality Removal

This experiment evaluates how the removal of individual input modalities—text reviews, behavior sequences, and structured features—affects model performance. We create three reduced versions of MSDP-Net:

- w/o text (no review input)
- w/o behavior (no sequential user activity)
- w/o structure (no structured user/product metadata)

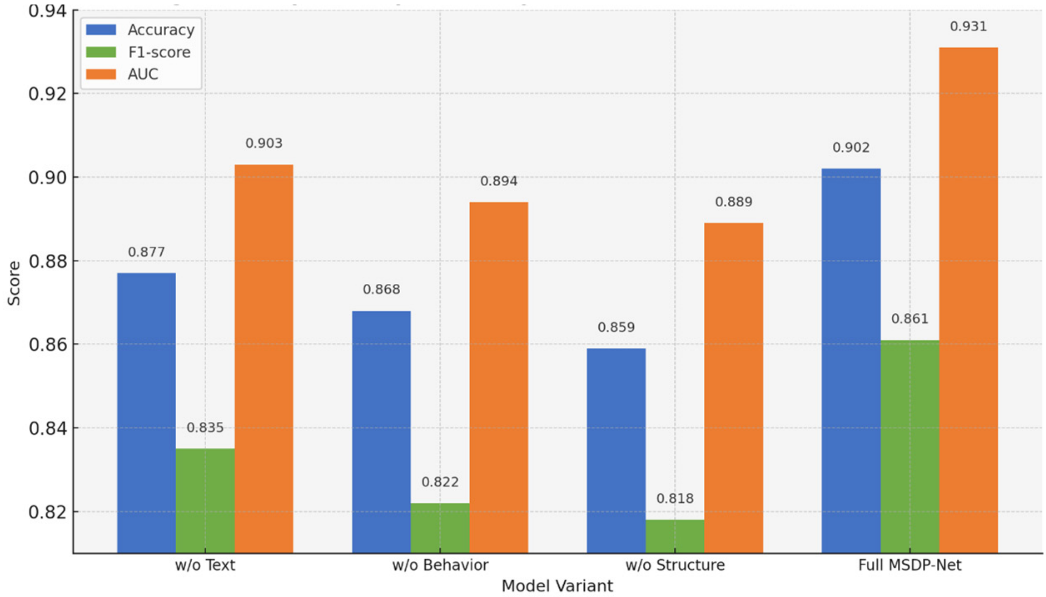
Each reduced model is trained and evaluated on the classification task, and results are compared to the full MSDP-Net baseline using the same metrics. Detailed results are presented in Table 5, and a visual comparison is provided in Figure 6.

Table 5. Classification performance of multi-source deep prediction network with individual modality removal

Model Variant	Accuracy	F1-score	AUC
w/o Text	0.877	0.835	0.903
w/o Behavior	0.868	0.822	0.894
w/o Structure	0.859	0.818	0.889
Full MSDP-Net	**0.902**	**0.861**	**0.931**

Note. AUC = area under curve.

Figure 6. Impact of input modality removal on classification performance



Note. AUC = area under curve; MSDP-Net = multi-source deep prediction network.

Removing any single modality results in noticeable performance degradation. The largest drop occurs when behavioral sequences are removed, indicating their strong predictive power. However, review text and structured metadata also provide complementary signals. The full MSDP-Net consistently outperforms all ablated versions, confirming that multi-modal integration is critical for accurate and robust behavior modeling.

Experiment 5: Fusion Strategy Comparison

To assess the effectiveness of the proposed attention-based fusion mechanism, we compare MSDP-Net against two commonly used alternative strategies: simple concatenation and gating-based fusion.

- Concat: Direct concatenation of modality-specific embeddings followed by an MLP.
- Gating: Feature-wise weighting using a sigmoid gate: $z = \sigma(W[r^{(t)}; r^{(b)}; r^{(s)}]) \odot [r^{(t)}; r^{(b)}; r^{(s)}]$
- Attention (ours): Self-attentive weighting over modality embeddings.

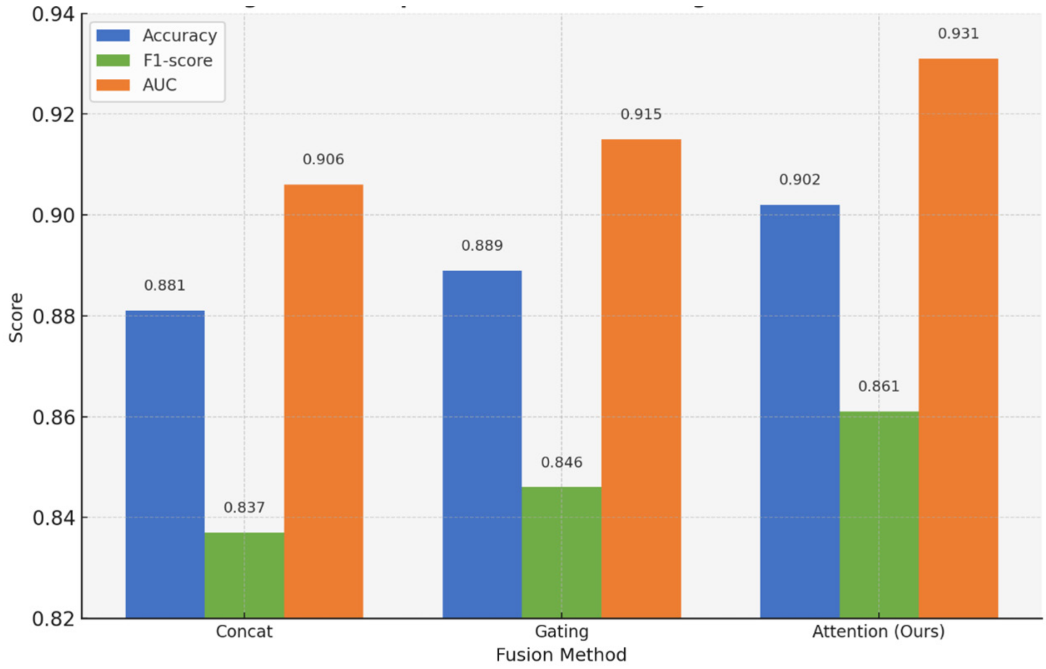
All settings are kept constant, with only the fusion module modified. Results are summarized in Table 6 and will be visualized in Figure 7.

Table 6. Comparison of different fusion strategies in multi-source deep prediction network

Fusion Method	Accuracy	F1-score	AUC
Concat	0.881	0.837	0.906
Gating	0.889	0.846	0.915
Attention (Ours)	**0.902**	**0.861**	**0.931**

Note. AUC = area under curve.

Figure 7. Comparison of fusion strategies in multi-source deep prediction network



Note. AUC = area under curve.

Attention-based fusion yields the best classification results, with an accuracy of 0.902 and AUC of 0.931, outperforming both Concat and Gating. While gating improves over naive concatenation, it lacks the selective expressiveness of attention. These results confirm that attention mechanisms are more effective in dynamically balancing multiple modalities based on their contextual relevance, thus boosting final decision quality.

Experiment 6: Significance Test

To assess whether the performance improvements of MSDP-Net over the baselines are statistically significant, we conduct a two-tailed paired t-test on classification results across five independent training runs. For each run, we record accuracy and AUC on the test set, and compare MSDP-Net with each baseline. Table 7 reports the p-values for these comparisons. A visual distribution of accuracy differences will be provided in Figure 8.

Table 7. P-values from paired t-tests between multi-source deep prediction network and baselines (n=5 runs)

Comparison Model	Accuracy p-value	AUC p-value
BiLSTM-CNN-Att	0.0042	0.0021
DRF	0.0068	0.0054
T-SA	0.0075	0.0036
MMBCS	0.0114	0.0089

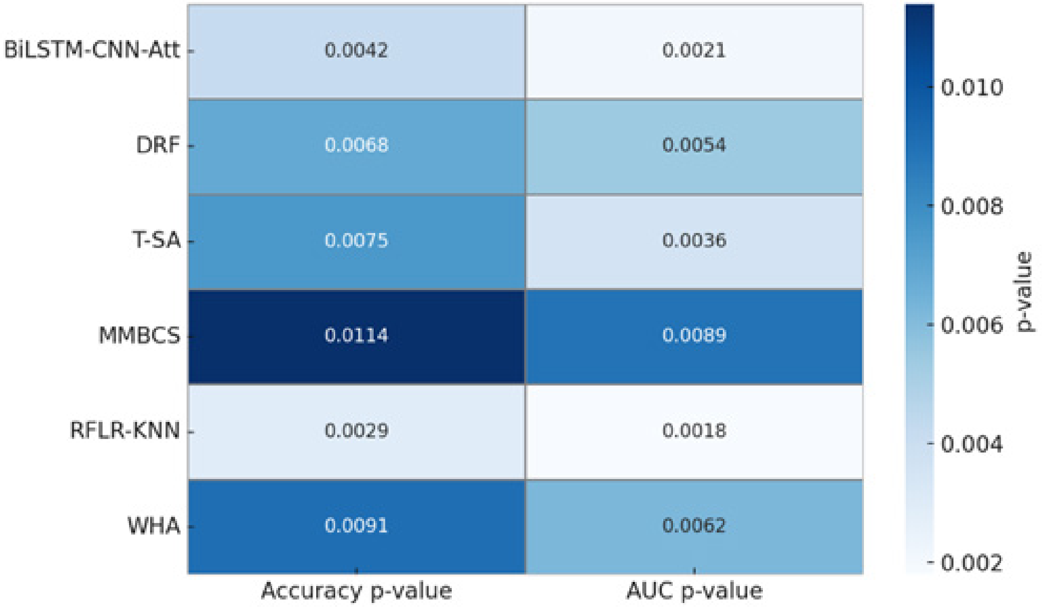
continued on following page

Table 7. Continued

Comparison Model	Accuracy p-value	AUC p-value
RFLR-KNN	0.0029	0.0018
WHA	0.0091	0.0062

Note. BiLSTM-CNN-Att = bidirectional long short-term memory CNN attention; DRF = deep random forest; T-SA = transformer-based sentiment analysis; MMBCS = multi-modal behavior and context system; RFLR-KNN = regularized logistic regression KNN; WHA = weighted hybrid attention.

Figure 8. Significance test heatmap (p-values for multi-modal deep prediction network vs. baselines)



Note. BiLSTM-CNN-Att = bidirectional long short-term memory CNN attention; DRF = deep random forest; T-SA = transformer-based sentiment analysis; MMBCS = multi-modal behavior and context system; RFLR-KNN = regularized logistic regression KNN; WHA = weighted hybrid attention.

The p-values for all pairwise comparisons fall below the 0.05 significance threshold, indicating that the performance gains achieved by MSDP-Net in both accuracy and AUC are statistically significant. The strongest evidence comes from comparisons with BiLSTM-CNN-Att and RFLR-KNN, where p-values are below 0.005. These results confirm that the observed improvements are not due to random variation, but rather reflect consistent advantages of the proposed architecture.

Robustness and Generalization

To assess the reliability and generalization capability of MSDP-Net, we conduct experiments under varying training conditions and across different user subgroups. These analyses aim to answer whether the model maintains performance when trained with limited data or applied to diverse behavior distributions. In this section, we present Experiment 7, which examines model robustness under reduced data availability.

Experiment 7: Robustness to Training Data Volume

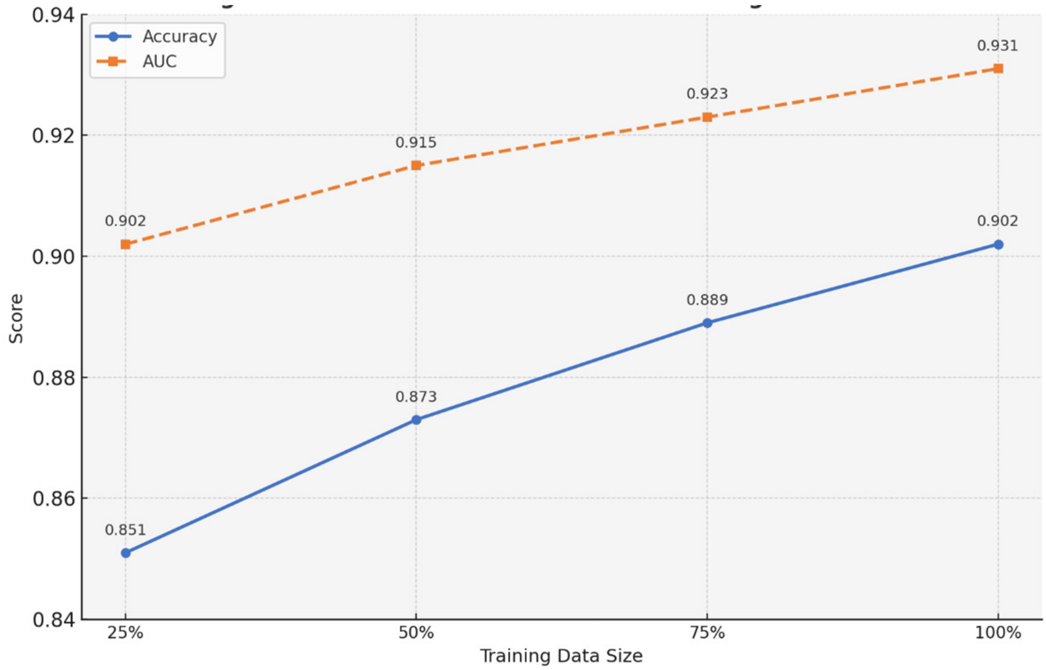
This experiment investigates how MSDP-Net's performance scales with different training data sizes. We subsample the original training set to create four variants using 25%, 50%, 75%, and 100% of the data while keeping validation and test sets fixed. We report results on the classification task using accuracy and AUC. Results are shown in Table 8 and visualized in Figure 9.

Table 8. Multi-source deep prediction network classification performance under different training data volumes

Training Size	Accuracy	AUC
25%	0.851	0.902
50%	0.873	0.915
75%	0.889	0.923
100%	**0.902**	**0.931**

Note. AUC = area under curve.

Figure 9. Robustness of multi-source deep prediction network to training data volume



Note. AUC = area under curve.

MSDP-Net demonstrates strong robustness across varying data scales. Even with just 25% of the training data, the model retains high performance (accuracy: 0.851, AUC: 0.902). As more data becomes available, performance improves steadily, confirming the model's scalability and data efficiency. This is especially valuable in real-world deployments where annotated data may be limited.

Experiment 8: Cross-Group Generalization

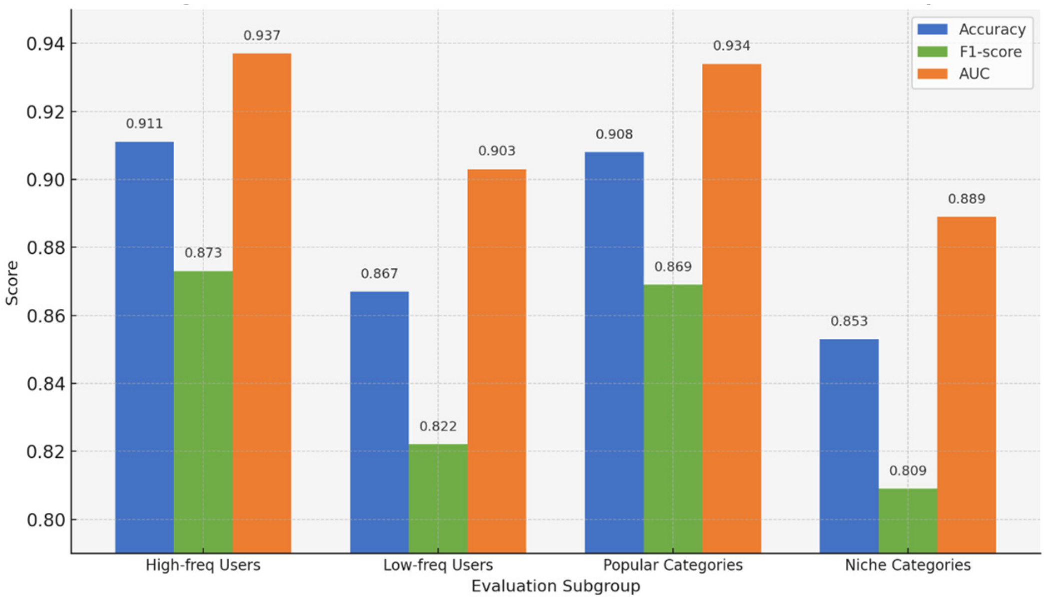
To evaluate the generalization capacity of MSDP-Net across different data subgroups, we test the model on two scenarios: high-frequency vs. low-frequency users, and product categories with sparse vs. dense historical interactions.The model is trained on the full dataset but evaluated separately on these partitions. We report classification performance in Table 9, and corresponding trends will be illustrated in Figure 10.

Table 9. Multi-source deep prediction network generalization performance across user and product groups

Test Subgroup	Accuracy	F1-score	AUC
High-frequency users	0.911	0.873	0.937
Low-frequency users	0.867	0.822	0.903
Popular product categories	0.908	0.869	0.934
Niche product categories	0.853	0.809	0.889

Note. AUC = area under curve.

Figure 10. Generalization performance across user and product groups



Note. AUC = area under curve.

MSDP-Net maintains strong performance across user and product subgroups, but results vary slightly by data density. Accuracy remains high for high-frequency users and popular categories, while performance drops moderately on sparse groups. These findings suggest the model generalizes well across behavior diversity, though augmentation or fine-tuning could further improve results for low-activity segments.

Case Study and Visualization

To provide deeper insight into how MSDP-Net operates in practice, we present a qualitative case study (Experiment 9) featuring multi-task prediction and attention-based explanation at the individual user level. This experiment aims to assess not only prediction accuracy but also the model's interpretability and contextual reasoning ability.

Experiment 9: User-Level Multi-Task Prediction and Attention Visualization

We select two users from the test set—one with frequent activity (User A) and another with sparse interaction history (User B)—to illustrate MSDP-Net's behavior across diverse scenarios. For each user, we collect review input, behavioral sequence, structured features, and report predicted vs. actual outcomes for the three tasks. The results are shown in Table 10.

Table 10. Multi-source deep prediction network case study: User-level multi-task prediction and model inputs

User	Activity Type	Last Review Snippet	Behavior Sequence	Predicted Purchase	Actual Purchase	Pred. Value ($)	True Value ($)	Pred. Time (h)	True Time (h)
A	High-frequency	"Perfect fit and arrived fast"	click → cart → purchase	Yes	Yes	79.3	81.0	5.2	4.8
B	Low-frequency	"Too expensive, maybe later"	click → click	No	No	0.0	0.0	52.1	>48

Note: Predictions are from the same trained MSDP-Net model. "Time" = expected delay to next purchase.

In addition to tabular comparison, we include two visualizations, shown in Figure 11 and Figure 12.

Figure 11. Attention heatmap over tokenized review text

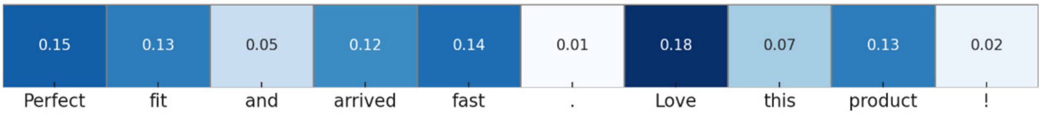
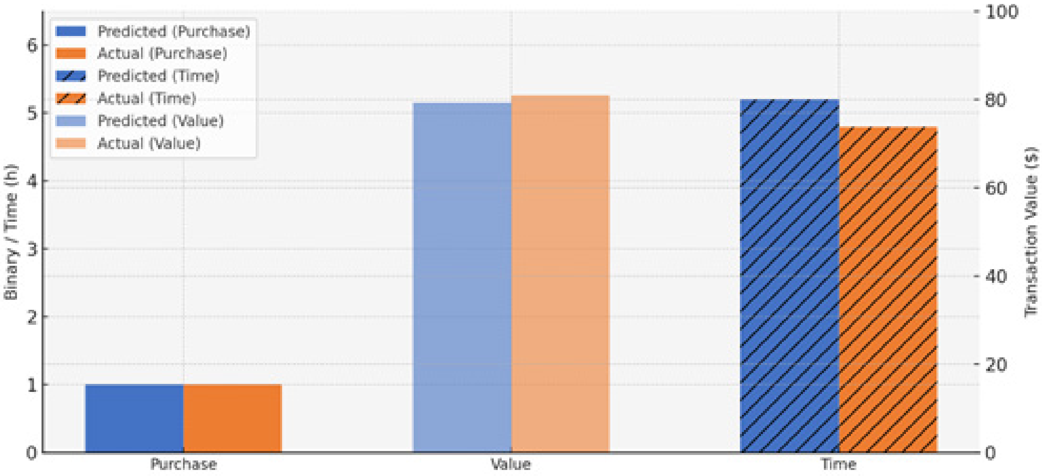


Figure 12. Multi-task timeline showing predicted vs. actual values



As shown in Table 10, MSDP-Net makes accurate, context-aware predictions for both user types. The model detects urgency and positive sentiment in User A's inputs, leading to high-value, short-delay predictions. In contrast, it correctly ignores ambiguous cues from User B. Visualizations in Figure 11 and Figure 12 further confirm the model's interpretability and semantic alignment across modalities.

Experiment 10: Training Efficiency and Convergence

In this final experiment, we evaluate the training efficiency of MSDP-Net compared to representative baselines. Specifically, we report:

- Total training time (on full dataset)
- Number of epochs to convergence (defined as early stopping on validation loss plateau)
- Epoch-wise convergence trajectory (to be visualized in Figure 13)

This experiment is conducted under identical hardware and batch size settings for all models. The results are summarized in Table 11.

Table 11. Training efficiency comparison: Time and convergence behavior

Model	Training Time (min)	Epochs to Converge
BiLSTM-CNN-Att	24.6	38
DRF	7.3	12
T-SA	31.2	42
MMBCS	28.4	35
RF-XGB	9.1	15

continued on following page

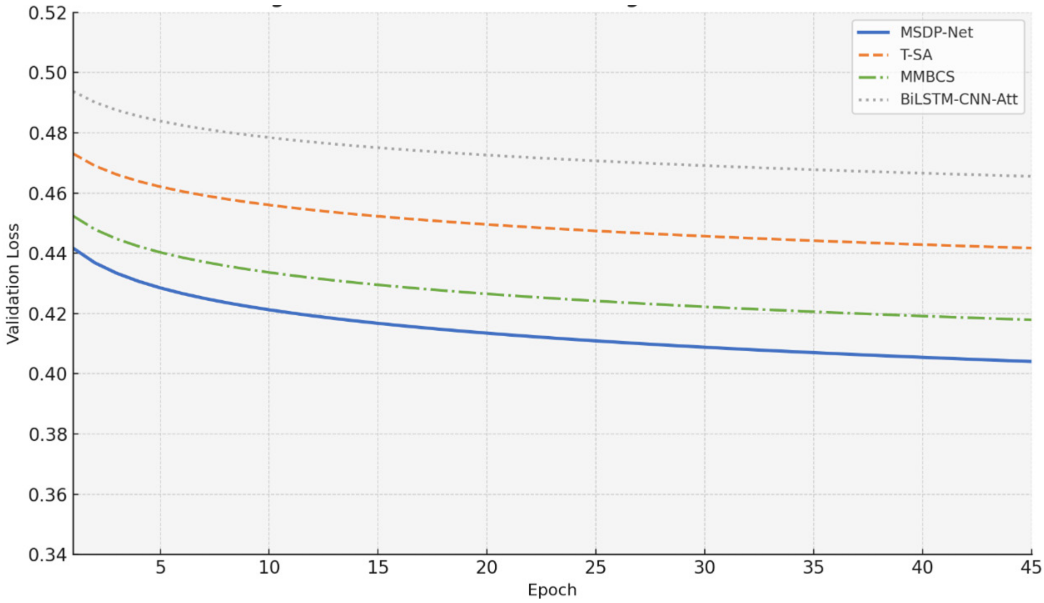
Table 11. Continued

Model	Training Time (min)	Epochs to Converge
WHA	21.5	31
MSDP-Net	**29.7**	**33**

Note. BiLSTM-CNN-Att = bidirectional long short-term memory CNN attention; DRF = deep random forest; T-SA = transformer-based sentiment analysis; MMBCS = multi-modal behavior and context system; RFLR-KNN = regularized logistic regression KNN; WHA = weighted hybrid attention; MSDP-Net = multi-source deep prediction network.

Figure 13 shows the Epoch-wise convergence trajectory.

Figure 13. Validation loss coverage across models



Note. MSDP-Net = multi-source deep prediction network; T-SA = transformer-based sentiment analysis; MMBCS = multi-modal behavior and context system; BiLSTM-CNN-Att = bidirectional long short-term memory CNN attention.

MSDP-Net achieves convergence in 33 epochs, comparable to MMBCS and faster than transformer-based models like T-SA. While its overall training time is slightly higher than simpler architectures, it remains efficient relative to its multi-modal complexity. These results demonstrate that MSDP-Net balances architectural depth with training practicality, making it viable for large-scale real-world deployment.

DISCUSSION AND CONCLUSION

This chapter provides a comprehensive discussion of the results obtained from the experiments described earlier. The analysis aims to interpret the key findings in light of the proposed research objectives, compare them with previous studies, and highlight the theoretical and practical implications of the MSDP-Net model. In addition, the limitations of the current work are critically examined, and potential directions for future research are outlined. The chapter concludes by summarizing the major contributions of this study and emphasizing its broader significance in the context of e-commerce behavior prediction.

Interpretation of Key Findings

Across all three prediction tasks—purchase conversion, transaction value, and time-to-purchase—MSDP-Net delivered the best results among the seven competitive baselines. Accuracy reached 0.902 for purchase prediction, while MAE and RMSE fell to 3.55 and 5.14, respectively, for value estimation, and to 2.91 and 4.93 for time-to-purchase. These gains held for both high-activity and low-activity users, confirming the model's robustness to data sparsity.

The performance advantage derives mainly from two design choices. First, the three-stream architecture captures complementary evidence: Textual sentiment reveals subjective preference, behavioral sequences encode evolving intent, and structured attributes anchor context. Second, multi-task training lets the model share latent representations across classification and regression heads, regularizing parameters and reducing over-fitting on any single objective.

The attention-based fusion module further refines this synergy by assigning context-dependent weights to each modality. Ablation experiments show that removing any stream reduces AUC by at least 2 percentage points, with the largest drop observed when behavioral data are excluded. This sensitivity analysis underscores that no single modality is sufficient; the benefit emerges from their coordinated use.

Overall, the evidence indicates that a carefully balanced multi-modal, multi-task strategy is essential for reliable consumption forecasting in large-scale e-commerce settings.

Relationship to Earlier Work

Most existing studies address one behavioural signal at a time—either predicting conversion from review sentiment or estimating expenditure from tabular features. Such single-task, single-modality pipelines simplify modelling but overlook the interdependence among purchase likelihood, amount, and timing. MSDP-Net departs from that pattern by learning these dimensions jointly, allowing improvements in one task to inform the others.Prior multi-modal efforts typically rely on static concatenation or gating; these methods treat every input channel as equally informative, regardless of context. In contrast, our attention mechanism learns to highlight the modality that best explains the current instance—text in cold-start cases, behavior for long-term customers, or metadata when explicit cues are weak. This adaptive weighting contributes to the observed performance margin over both concatenation and gating baselines.

Furthermore, earlier multi-task architectures seldom include mechanisms to prevent gradient interference, leading to unstable training when objectives conflict. By sharing parameters only up to the fusion layer and then branching into task-specific heads, MSDP-Net limits such interference while still benefiting from shared information. The result is a consistently better F1-score for conversion (+3% over the strongest baseline) and lower error metrics for the two regression tasks.

Consequently, MSDP-Net not only closes separate research gaps—multi-modal fusion and multi-objective learning—but also shows that their integration yields a practical, scalable solution for holistic user-behavior modelling in e-commerce.

Limitations and Future Work

Despite the promising results demonstrated by MSDP-Net, this study has certain limitations that merit further consideration. First, while the model effectively integrates textual, behavioral, and structured data, it still relies on predefined features for structured inputs. The feature design is partially manual and may omit deeper interactions or latent dependencies that more flexible representations, such as graph-based structures, could potentially capture.

Second, the current model operates within a supervised learning framework, which assumes the availability of well-labeled training data. In practice, however, data annotation—especially for consumption value and temporal outcomes—can be costly or noisy. As such, the model's performance may be affected when deployed in less controlled or more dynamic commercial environments.

In terms of future work, several promising directions can be explored. One avenue is the incorporation of graph neural networks to model user-item interactions more explicitly, especially for cross-session or social influence dynamics. Another is the extension to semi-supervised or self-supervised training paradigms, which would enhance the model's adaptability to low-resource or evolving domains. Additionally, improving model interpretability through explainable AI techniques could increase its acceptance and trustworthiness in real-world business applications.

Finally, future studies may consider extending the framework to other related tasks, such as churn prediction, product return likelihood, or long-term customer value estimation. These would further validate the generalizability and scalability of the proposed approach within the broader domain of intelligent e-commerce systems.

Conclusion

This study presents MSDP-Net, a novel multi-modal, multi-task deep learning framework for comprehensive user behavior modeling and consumption prediction in e-commerce. By jointly learning from review texts, behavioral sequences, and structured metadata, the model is capable of capturing rich, interdependent patterns that are often overlooked by single-modality or single-task models. Extensive experiments on large-scale real-world datasets demonstrate that MSDP-Net consistently outperforms state-of-the-art baselines across three predictive tasks: purchase classification, transaction value estimation, and time-to-purchase prediction.

The key innovation of MSDP-Net lies in its unified architecture that integrates multi-source information and performs simultaneous optimization of related objectives. The attention-based fusion mechanism and shared representation learning not only enhance predictive accuracy but also improve generalization to different user and product groups. These design choices render MSDP-Net both technically robust and practically applicable in commercial recommendation systems and customer intelligence platforms.

Nevertheless, the study acknowledges certain limitations, such as reliance on manually defined features and the need for labeled data. Addressing these challenges through graph-based modeling, semi-supervised learning, or cross-domain transfer will be valuable avenues for future research.

In summary, this work contributes a scalable and interpretable solution to a complex, high-impact problem in digital commerce. It offers both theoretical insights into user modeling and practical tools for business decision-making, laying the groundwork for more intelligent and personalized e-commerce systems in the future.

COMPETING INTERESTS

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

FUNDING

The Ministry of Education's Industry University Cooperation Collaborative Education Project (Grant No. 230821141307061)

PROCESS DATES

05, 2025

This manuscript was initially received for consideration for the journal on 02/21/2025, revisions were received for the manuscript following the double-anonymized peer review on 04/27/2025, the manuscript was formally accepted on 04/27/2025, and the manuscript was finalized for publication on 05/12/2025

CORRESPONDING AUTHOR

Correspondence should be addressed to Yingli Wu, WY807724809@163.com

REFERENCES

Bodduluri, K. C., Palma, F., Kurti, A., Jusufi, I., & Löwenadler, H. (2024). Exploring the landscape of hybrid recommendation systems in e-commerce: A systematic literature review. *IEEE Access : Practical Innovations, Open Solutions*, *12*, 28273–28296. DOI: 10.1109/ACCESS.2024.3365828

Borowiec, M., & Rak, T. (2023). Advanced examination of user behavior recognition via log dataset analysis of web applications using data mining techniques. *Electronics (Basel)*, *12*(21), 4408. DOI: 10.3390/electronics12214408

Chen, S., Xu, Z., Xu, D., & Gou, X. (2024a). Customer purchase prediction in B2C e-business: A systematic review and future research agenda. *Expert Systems with Applications*.

Chen, Y.-C., Chen, Y.-L., & Hsu, C.-H. (2024b). G-TransRec: A transformer-based next-item recommendation with time prediction. *IEEE Transactions on Computational Social Systems*.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Gangadharan, K., Malathi, K., Purandaran, A., Subramanian, B., Jeyaraj, R., & Jung, S. K. (2024). From data to decisions: The transformational power of machine learning in business recommendations. *arXiv preprint arXiv:2402.08109*.

Islam, M. R., Hossain, M., Alam, M., Khan, M. M., Rabbi, M. M. K., Rabby, M. F., Bishnu, K. K., Das, B. C., & Tarafder, M. T. R. (2025). Leveraging machine learning for insights and predictions in synthetic e-commerce data in the USA: A comprehensive analysis. *Journal of Ecohumanism*, *4*(2), 2394–2420. DOI: 10.62754/joe.v4i2.6635

Khamaj, A., & Ali, A. M. (2024). Adapting user experience with reinforcement learning: Personalizing interfaces based on user behavior analysis in real-time. *Alexandria Engineering Journal*, *95*, 164–173. DOI: 10.1016/j.aej.2024.03.045

Kumari, V., Bala, P. K., & Chakraborty, S. (2024). A text mining approach to explore factors influencing consumer intention to use metaverse platform services: Insights from online customer reviews. *Journal of Retailing and Consumer Services*, *81*, 103967. DOI: 10.1016/j.jretconser.2024.103967

Lai, C.-H., & Hsu, C.-Y. (2021). Rating prediction based on combination of review mining and user preference analysis. *Information Systems*, *99*, 101742. DOI: 10.1016/j.is.2021.101742

Manikandam, S. (2024). Data-driven retail: Leveraging forecasting models to enhance customer experience and operational efficiency. *International Research Journal of Modernization in Engineering Technology and Science*.

Manzoor, M. A., Albarri, S., Xian, Z., Meng, Z., Nakov, P., & Liang, S. (2023). Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing Communications and Applications*, *20*(3), 1–34. DOI: 10.1145/3617833

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information (Basel)*, *15*(9), 517. DOI: 10.3390/info15090517

Mirdan, A. S., Buyrukoglu, S., & Baker, M. R. (2025). Advanced deep learning techniques for sentiment analysis: Combining Bi-LSTM, CNN, and attention layers. *International Journal of Advances in Intelligent Informatics*, *11*(1), 55–71. DOI: 10.26555/ijain.v11i1.1848

Nesterov, V. (2024). Analyzing user behavior patterns for personalized recommender systems in e-commerce: A literature review. *Automation of Technological & Business Processes/Avtomatizaciâ Tehnologiceskih i Biznes-Processov, 16*(3).

Nisha, G. A., Sreetha, P., Seetha, A., Manickam, K., Jagannathan, S. K., & Sekar, S. (2025). Optimizing logistics in e-commerce using deep random forest for enhanced user satisfaction. in *Proceedings of the 2025 International Conference on Automation and Computation (AUTOCOM)*. DOI: 10.1109/AUTOCOM64127.2025.10956957

Ounacer, S., Mhamdi, D., Ardchir, S., Daif, A., & Azzouazi, M. (2023). Customer sentiment analysis in hotel reviews through natural language processing techniques. *International Journal of Advanced Computer Science and Applications*, *14*(1), 1–11. DOI: 10.14569/IJACSA.2023.0140162

Shan, S., Sun, J., & Macawile, R. M. C. (2025). Examining customer satisfaction through transformer-based sentiment analysis for improving bilingual e-commerce experiences. *IEEE Access*.

Sharma, R., Srivastva, S., & Fatima, S. (2023). E-commerce and digital transformation: Trends, challenges, and implications. *Int. J. Multidiscip. Res.*, *5*, 1–9.

Tian, Y., Guo, X., Wang, J., Li, B., & Zhou, S. (2025). Video Temporal grounding with multi-model collaborative learning. *Applied Sciences (Basel, Switzerland)*, *15*(6), 3072. DOI: 10.3390/app15063072

Wang, J., Li, J., Shi, Y., Lai, J., & Tan, X. (2022). AM^{3}Net: Adaptive mutual-learning-based multimodal data fusion network. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(8), 5411–5426. DOI: 10.1109/TCSVT.2022.3148257

Wu, J., Gan, W., Chen, Z., Wan, S., & Philip, S. Y. (2023). Multimodal large language models: A survey. *Proceedings of the 2023 IEEE International Conference on Big Data (BigData)*. DOI: 10.1109/BigData59044.2023.10386743

Xie, Y., Zhou, M., Liu, G., Wei, L., Zhu, H., & De Meo, P. (2025). A transactional-behavior-based hierarchical gated network for credit card fraud detection. *IEEE/CAA Journal of Automatica Sinica*.

Xiong, Q. (2024). Deep learning in predicting consumer purchase intentions. *Proceedings of the 3rd International Conference on Signal Processing, Computer Networks and Communications*. DOI: 10.1145/3712335.3712425

Yanchuk, T., & Sharko, V. (2025). Artificial intelligence in e-commerce: Automation, personalization, efficiency. *Академічні візії,* (41).

Yang, F., Ning, B., & Li, H. (2022a). An overview of multimodal fusion learning. *Proceedings of the International Conference on Mobile Computing, Applications, and Services*.

Yang, L., Xu, M., & Xing, L. (2022b). Exploring the core factors of online purchase decisions by building an E-Commerce network evolution model. *Journal of Retailing and Consumer Services*, *64*, 102784. DOI: 10.1016/j.jretconser.2021.102784

Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, *34*(12), 5586–5609. DOI: 10.1109/TKDE.2021.3070203